Tech., Inst., Cognition and Learning, Vol. 2, pp.103-140
Reprints available directly from the publisher
Photocopying permitted by license only

Using Automated Questions to Assess Reading Comprehension, Vocabulary, and Effects of Tutorial Interventions

JACK MOSTOW*, JOSEPH BECK, JULIET BEY[1] , ANDREW CUNEO, JUNE SISON, BRIAN TOBIN[2] , & JOSEPH VALERI

Project LISTEN, RI-NSH 4213, 5000 Forbes Avenue
Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA
http://www.cs.cmu.edu/~listen

We describe the automated generation and use of 69,326 comprehension cloze questions and 5,668 vocabulary matching questions in the 2001-2002 version of Project LISTEN's Reading Tutor used by 364 students in grades 1-9 at seven schools. To validate our methods, we used students' performance on these multiple-choice questions to predict their scores on the Woodcock Reading Mastery Test. A model based on students' cloze performance predicted their Passage Comprehension scores with correlation R = .85. The percentage of vocabulary words that students matched correctly to their definitions predicted their Word Comprehension scores with correlation R = .61. We used both types of questions in a within-subject automated experiment to compare four ways to preview new vocabulary before a story – defining the word, giving a synonym, asking about the word, and doing nothing. Outcomes included comprehension as measured by performance on multiple-choice cloze questions during the story, and vocabulary as measured by matching words to their definitions in a posttest after the story. A synonym or short definition significantly improved posttest performance compared to just encountering the word in the story – but only for words students didn't already know, and only if they had a grade 4 or better vocabulary. Such a preview significantly improved performance during the story on cloze questions involving the previewed word – but only for students with a grade 1-3 vocabulary.

Keywords: Assessment, external validity, reading comprehension, vocabulary, multiple-choice cloze tests, embedded experiment, within-subject experiment, curriculum-based assessment, validation, Woodcock Reading Mastery Test, Project LISTEN, Reading Tutor.

*Corresponding author: mostow@cs.cmu.edu
[1]Now at University of Southern California Law School, Los Angeles, CA 90089.
[2]Now at Office of Information Resources and Technology, Stanford University School of Medicine, Alway Building, M-121, 300 Pasteur Drive, Stanford, CA 94305-5119, USA.

INTRODUCTION

We report on automatically generating questions to estimate children's reading comprehension and vocabulary for two purposes. One purpose is to assess the student, and is validated by predicting test scores. The other purpose is to evaluate the effects of tutorial interventions, and is validated by demonstrating significant differences between their effects on moment-to-moment variations in comprehension.

Assessing an educational experience requires fitting the assessment to the experience. For example, an assessment of how well a reader comprehended a particular story must be specific to that story. An educational assessment should also be validated. For example, questions about a particular story may or may not provide a valid assessment of how well the reader comprehended it. However, validating assessments against accepted measures is costly in time, effort, and students. Validating assessments that are customized to an individual student's educational experience would therefore seem at best exorbitant and at worst impossible. This paper offers a way out by automating the generation of assessment questions, and validating the automated *process* by correlating its assessments against accepted measures. The resulting validation applies not only to a particular set of items, but also to any future assessments generated by the same process. We illustrate and validate such a process in the context of an automated reading tutor.

Project LISTEN's Reading Tutor listens to children read (Mostow & Aist, 2001), and helps them learn to read (Mostow *et al.*, 2002a; Mostow *et al.*, 2003a; Mostow *et al.*, 2004, in press). To balance learner control with tutorial guidance, the 2001-2002 version of the Reading Tutor used in this study took turns with the student to pick from its hundreds of stories, and used students' assisted reading rate to adjust the story level it picked (Aist & Mostow, 2004, in press), ranging from kindergarten to grade 7 (disguised as K, A, B, ..., G to avoid embarrassing poor readers). Thus every student read 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 – and evaluate how specific tutorial interventions affect them? We assess comprehension, vocabulary, and effects of vocabulary previews in subsequent main sections of this paper.

AUTOMATED COMPREHENSION QUESTIONS

Existing assessments of children's vocabulary and comprehension such as the Woodcock Reading Mastery Test (WRMT) (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 as yet too inaccurate on such unpredictable speech.

Researchers and educators generate *cloze* tests from a given text by replacing one or more words with blanks to fill in. The task of inferring the closure of the text (that is, reconstructing the omitted portion of the text) tests the reader's comprehension. Vacca *et al.* (1991, pp. 270-272) describe several variants of cloze questions and point out that "cloze-type materials are available commercially from publishing companies. However, teachers often produce the most effective cloze passages, because they are in the best position to gear the material to the needs of their students." The cloze method has been applied not only to comprehension of natural language, but also to the Pascal programming language to measure students' comprehension of computer programs (Entin, 1984).

The mechanical nature of cloze test generation is conducive to automation, as in this freely available software (Drott, n.d.):

> *A 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.*

In this section we present a fully automated approach to generating cloze questions, scoring them, and using them to assess students' comprehension and vocabulary.

In using cloze items to assess vocabulary and comprehension, we had to decide 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 would test what students retained from the story, but would conflate comprehension with memory of specific words, suffer from recency effects, and feel like a test. We decided to insert occasional cloze questions in a story just before displaying a sentence to read. This decision offers the additional advantage of assessing comprehension where it occurs – while reading the text – at the possible risk of disrupting comprehension in order to measure it.

Most words in English text (apart from highly predictable function words) are too hard to guess from context (Beck *et al.*, 2002). To overcome this unpredictability problem, we decided to make our cloze questions be multiple-choice instead of fill-in-the-blank. This tactic traded one problem for another. We no longer had to score arbitrary student responses, but besides choosing which words to turn into cloze items, we also had to generate appropriate distractors for them. Before describing our methods to address these and other implementation issues, we illustrate the cloze questions they produced.

Example

Our example begins as a student started to read Aesop's fable, "*The Ant and the Grasshopper*." As Figure 1 shows, the Reading Tutor displayed the first sentence for the student to read aloud: *In a field one summer's day a Grasshopper was hopping about, chirping and singing to its heart's content.* When the Reading Tutor detected a mistake, hesitation, or request for help, it gave spoken assistance, such as highlighting the word *hopping* and reading it aloud.

After a few sentences, the Reading Tutor inserted a multiple-choice cloze question, as Figure 2 shows. The Reading Tutor said "click on the missing word," and read aloud: *"I am helping to lay up food for the winter," said the Ant, "and____ you to do the same." bother; recommend; chat; grasshopper.* It displayed the cloze prompt at the top of the screen and read each choice aloud, highlighting its background in yellow as it did so. If the student had not clicked on a choice yet, the Reading Tutor read the list again.

Item difficulty varied with several factors discussed below. What did this particular item test? Three of the choices happened to be verbs, including the correct answer *recommend*. Information about part of speech could rule out *grasshopper*, improving the odds of guessing correctly to 1 in 3 – given the

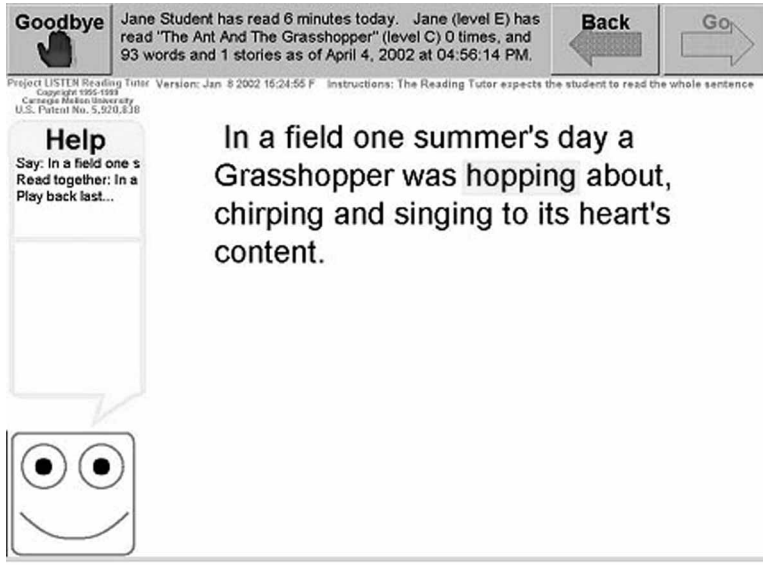

FIGURE 1
Student starts reading a story.

linguistic knowledge and metacognitive skills (either of which young readers might lack) to detect and eliminate inconsistent choices, and guess from the remaining ones. Additional knowledge, such as semantics, was required to distinguish among the remaining choices.

When the student clicked on a choice, the Reading Tutor did not give explicit feedback, but went on to display the complete sentence for the student to read, implicitly indicating the correct answer. Thanks to having just heard the Reading Tutor read the cloze item, the student presumably read the sentence faster than she would have on her own. Consequently, the net time cost of inserting the cloze question was ostensibly less than the time it took to administer, and might even have been negative for a student who would otherwise have read the sentence very slowly. 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. Possibly the cloze activity itself built comprehension skills by exercising them, 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

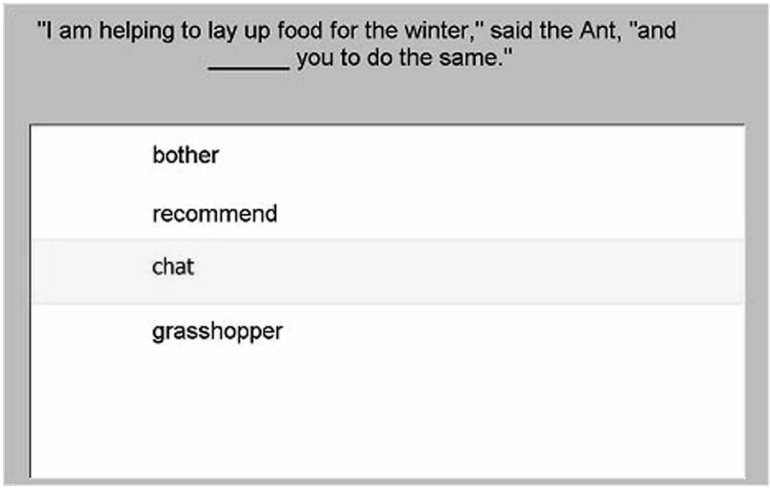

FIGURE 2
Example of a "hard word" cloze question.

information integration suggested that "cloze scores may reflect inter-sentential comprehension sufficiently to warrant their continued use in assessment," based on a study of 281 fifth graders (McKenna & Layton, 1990). A cloze question that exercised such integration was *"Why bother about _____?" food; winter; dying; passed.* Although the words were fairly easy, the question was challenging because only one choice (*passed*) could be ruled out based on its part of speech or other information local to the sentence. The correct choice (*winter*) depended on the preceding context. *Food* or *dying* might be reasonable, but not what the author wrote.

Considerations in Generating Cloze Questions

We now discuss some of the design considerations illustrated by the example.

Text difficulty: The difficulty of a cloze question is determined in part by the story in which the clozed sentence occurred: harder stories tend to have longer, more complex sentences, with harder words.

Vocabulary: The familiarity of the target word and distractors affects the difficulty of a cloze question. Matching them by word frequency (Coniam, 1997) both adjusts difficulty and avoids bias based on familiarity. 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:

1. **Sight words:** the most frequent 225 words in our table, approximately the same as the "Dolch list" (Dolch, 1936). These words cover over half the word tokens in school text, and hence are emphasized in early reading.
2. **Easy words:** the most frequent 3000 words (a heuristic cutoff) in our tableexcluding the top 225.
3. **Hard words:** all 25,000 words in our frequency table except for the to 3000.
4. **Defined words:** Story words explicitly annotated as warranting explation to the student. This category was not defined in terms of frequency,so it overlapped with the other ranges. The Reading Tutor explained only some of the defined words, for a separate experiment (see *Evaluation of Vocabulary Previews* section) that tested different ways to preview new words.

Similarity: The relationship of the distractors to the correct target word also affects the difficulty of the question. For example, if they have the same part of speech, then the item requires semantic processing to ensure a correct answer. Coniam (1997) matched distractors both by frequency and by word class using an automatic part of speech tagger. The word tagger's 5% error rate caused some of the low quality items generated by his system.

Modality: The Reading Tutor gave students reading assistance on difficult words and sentences. To avoid frustrating them by presenting a cloze question without such assistance, we decided to have the Reading Tutor read the cloze question aloud by playing back the already-recorded human narration of the sentence, minus the deleted word. This decision also reflected our desire to assess comprehension of the text rather than ability to decode the question. The Gray Oral Reading Test (Wiederholt & Bryant, 1992) likewise uses multiple comprehension questions read aloud by the examiner, many of them cloze questions.

We chose distractors from the story, rather than from a general lexicon, for the following three reasons:

1. **Voice matching**: The Reading Tutor used digitized human speech rather than a synthesizer. 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.

2. **Word recency**: Students might be biased toward picking words they had encountered recently, in particular earlier in the story. Choosing distractors from the story rather than from a general lexicon made them as likely as the target word to have appeared recently.

3. **Social acceptability**: Words chosen from an unrestricted lexicon might be offensive. Choosing words from the story meant that they had already been judged acceptable by whichever adult added the story.

Table 1 gives additional examples of cloze items, per word type. Only levels C and higher (grade 3 and above) had "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 tended to increase with story level. The number of distractors with the same part of speech as the target word ranged from 0 to 3, due to random variation. Distractors were often semantically related to the target word (e.g. *beak*, *hens*), thanks to selecting them from the same story instead of from a general lexicon.

Automated generation of multiple-choice cloze questions

We now describe how the 2001-2002 Reading Tutor generated and presented cloze questions, starting with how it decided which category of cloze question to insert when. The Reading Tutor had an event-driven control architecture. For example, an ev_before_new_sentence event fired before each sentence of a story. Possible responses to that event were the four categories of cloze questions, plus "Do Nothing" (which simply went on to display the new sentence).

The Reading Tutor chose probabilistically from these responses, with a weight of 10,000 for "defined" words, a weight of 100 for "easy," "hard," and "sight" words, and a weight of 400 for "Do Nothing." The effect of the high weight for the "defined" words category was to prefer it at almost every opportunity – that is, whenever at least one word in the sentence and least three other words in the story were marked as defined. If the chosen response failed – for example, if no word in the sentence was marked as defined – then the Reading Tutor picked another response probabilistically until it found one it could fire, possibly Do Nothing.

We represented responses in a language we developed to express activities in a concise form that we could understand and the Reading Tutor could execute. The response code for each word type was almost identical, and less than 20 lines long. We describe some code for precision – it's the representation we ourselves consult when in doubt – and to convey the

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

Word Type Level	Story	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

generality of the specification, and the conciseness afforded by some key constructs of the activity language.

One such construct provided fast filtered random selection without replacement. Thus the following statement randomly selected from the sentence a word whose frequency put it in the top 3000 words but not the top 225:

```
Set_variable test_word a_sentence_word
_WHERE_ (WordRank(test_word) < 3001)
_WHERE_ (WordRank(test_word) > 225)
```

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

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

To present a cloze item, the Reading Tutor used its generic "talking menu" mechanism for multiple-choice questions. This mechanism displayed and spoke a prompt and a list of possible choices to click on. For cloze items, the Reading Tutor first said, "Click on the missing word." Next it read the sentence aloud, minus the test word, by playing the appropriate portions of the recorded sentence narration. Then it read aloud the displayed menu of choices, listed in randomized order, and consisting of the test word and the three distractors. It logged the test word, the distractors, the cloze sentence, the student's answer, and additional context information such as timestamp and student ID.

Data set for cloze responses

How accurately could we assess students' comprehension using automatically generated cloze questions? To answer this question, we analyzed items from 99 Reading Tutors used in eight schools over the 2001-2002 school year. A perl script on each Reading Tutor sent logged data back each night by ftp to Carnegie Mellon, where it was parsed into a suitable form to import into a statistical analysis package (SPSS, 2000).

The following analysis is restricted to 364 students individually pretested

on the Woodcock Reading Mastery Test (Woodcock, 1998) before they used the Reading Tutor. These 364 students were 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 came 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 varied 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 items (0.5%) by clicking *Back*, and 265 items (0.4%) by waiting long enough for the Reading Tutor to time out, but they answered the remaining 67,971 items (98.0%).

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 was 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 had read before, the Reading Tutor inserted cloze items only if the student had 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

TABLE 2
Per-student number of cloze items and percent correct, by grade and overall

	Grade:	1	2	3	4	5	6	7	9	All
number students		35	78	47	72	35	29	17	2	315
number 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
percent 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%

prompt more than once. 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.

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

The 67,971 responses analyzed included 17,566 "sight word" items, 17,092 "easy word" items, 12,010 "hard word" items, and 21,303 "defined word" items. 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% correct for the 3,163 level K items, up to 69% for the 6,031 level C items, then down to 55% for the 7,061 level E items, and finally back up to 59% for 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 20 or more responses; they spanned grades 1-7, plus two 9th graders. 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, which only 260 of 315 students were tested on.

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 the course of the year. 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.

Hasty Responses

How much guessing was there? We don't know how to tell in general, but we can distinguish the important case of "hasty 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 word type, from 34% for sight words down to 27% for defined words, suggesting that most but not all of them were thoughtless guesses. In contrast, thought*ful* guesses based on considering and eliminating choices should do better.

As one might expect, some students responded hastily more often than others. The per-student rate of "hasty responses" averaged 3.9% overall, but 1% or less for over half of the students. Hasty responses rose 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. The 2002-2003 version of the Reading Tutor gave explicit praise for correct cloze responses, as an incentive to reward thoughtful responses.

Reliability

The reliability of a measure characterizes the consistency of its results. How well does a student's performance on one half of an m-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, i.e., pairing the leftover items despite mismatches.

We used SPSS to compute the Guttman Split-half test of reliability. The resulting coefficient depended on the minimum permissible number of items m, and ranged from .83 for $m \geq 10$ (338 students) to .95 for $m \geq 80$ (199 students). Thus student performance on a sufficient number of cloze items was indeed highly reliable.

External validity

Did 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 ($p < .001$) with WRMT Passage Comprehension Raw Score (R = .51), WRMT Word Comprehension Weighted Score (R = .53), and GORT (R = .40), but much less than these tests did with each other: these two WRMT subtests each correlated 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 as a vector of 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. Defined words started at level C, so the 4 question types and 8 story levels made 29 such combinations (3 each for levels K, A, and B, plus 4 each for levels C through G). We used backward regression in SPSS to build a separate model for each test by regressing test score against all the 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.

To estimate the performance of each model on unseen data from a similar distribution of students, we used the "leave-1-out" method. This method, standard in machine learning, adjusts the prediction for each student by training the same model without that student's data, and correlates these adjusted predictions against the actual scores. The resulting correlation measures how well the model generalizes to students drawn from a similar distribution.

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 first graders can read independently), and highest in grade 5. Predictive validity was higher for Word Comprehension and Passage Comprehension than for Word Attack and Word Identification, both overall and within almost every grade, suggesting that cloze performance

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**	**Leave 1 out**
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	0.69
Word Identification	0.21	0.65	0.64	0.62	0.85	0.57	0.40	1.00	0.84	0.82
Word Comprehension	0.44	0.73	0.65	0.71	0.85	0.71	0.55	1.00	0.86	0.84
Passage Comprehension	0.17	0.75	0.67	0.73	0.87	0.59	0.40	-1.00	0.85	0.83
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

was not just measuring overall reading proficiency. Predictive validity for GORT was even higher in grades 1 and 6, but lower in grades 2-5.

Figure 3 plots predicted scores (adjusted by the leave-1-out method) against actual scores, showing students as digits from 1 to 9 according to what grade they were in.

Construct validity

What skills did the cloze items actually test? The Reading Tutor read 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 might measure listening comprehension, or at least comprehension of "assisted reading," whereas WRMT measures independent reading skills. It might be interesting to see if silent presentation of items 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 tested the ability to decide whether a given word fits in a particular context, whereas a fill-in format tests the ability to predict the

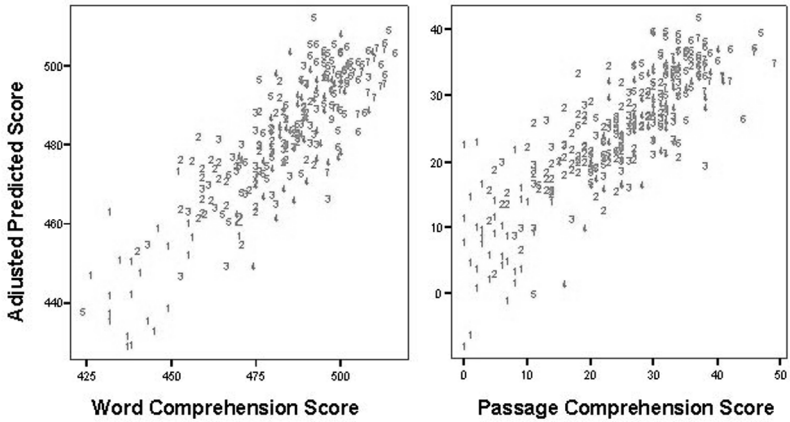

FIGURE 3
Scatterplots of adjusted predicted scores by actual WRMT scores of 315 students in grades 1-9

word outright. The multiple-choice format was arguably more valid for testing readers' metacognitive ability to judge whether they had identified a word correctly by seeing if it made sense in context. Inserting questions throughout the reading of a story, rather than at the end, was arguably more valid for testing comprehension processes that occurred while students were reading, but it did 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 required 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 varied randomly in the degree to which they required semantics or intersentential context to answer. Items with choices deliberately matched – or mismatched – by part of speech might help tease these skills apart.

Finally, it is important to acknowledge that our model made predictions based not simply on the percentage of items correct in each category, but also on the amount and levels of material read. Using the amount of reading as a predictive feature exploited the extent to which better readers read more, but hurt predictive accuracy to the extent that this amount was affected by other factors such as differences among teachers. Using the distribution of grade levels at which the student and Reading Tutor chose

stories exploited the extent to which they correlated with the student's actual level of proficiency.

Relation to other work on automated generation of cloze questions

How does this research relate to previous 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 to about 60 twelfth grade ESL students in Hong Kong. He rated the methods by percentage of "acceptable" questions, defined by "a facility index of 30%-80% and a discrimination index greater than 0.2" (p. 23). Picking test words by frequency range or word class yielded higher percentages of acceptable items than picking every n^{th} word, which picked too many high-frequency words.

The work presented here differs in several respects. Our data came 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 filtered by hand and administered to all students. We chose distractors from the same story, not from a large general lexicon. We matched distractors by gross frequency range also, but not by word class. The Reading Tutor presented cloze items aloud, not silently. Finally, we evaluated correlation to established measures of comprehension and vocabulary, not percentage of acceptable test items.

AUTOMATED VOCABULARY QUESTIONS

Besides assessing comprehension of text passages, we wanted to test students' comprehension of individual words in a story, both before and after reading the story, for an experiment to be described later.

Generation of vocabulary questions

To test the student's ability to match words to their meanings, the

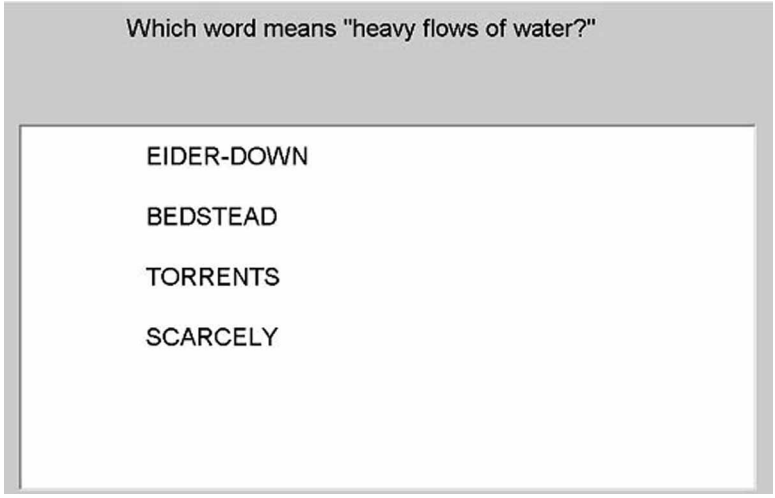

FIGURE 4
Example of a multiple-choice question to match word meaning.

Reading Tutor used multiple-choice questions presented in the same way as the cloze questions described earlier. The meaning of a word might differ from one story to another, but seldom within the same story. Therefore, the Reading Tutor associated word definitions with stories, rather than with specific instances of a word in a story.

The prompt for each question displayed and read aloud a short, story-specific definition written according to the guidelines in (McKeown, 1993). The choices were four vocabulary words from the same story, as Figure 4 illustrates. Successive questions highlighted and read the same four words aloud, re-randomizing their order:

- *Which word means "heavy flows of water?" eider-down; bedstead; torrents; carcely.*
- *Which word means "duck feathers?" eider-down; bedstead; scarcely; torrents*.
- …

The Reading Tutor selected these four words at random from story words hand-annotated as vocabulary likely to merit explanation (Aist, 2000). Only stories at and above level C contained such words, because earlier work (Aist, 2002; Beck *et al.*,

2002) suggested that children below a third grade level of reading proficiency were unlikely to learn vocabulary by reading. The Reading Tutor generated vocabulary questions only for stories with at least four annotated vocabulary words.

Data set for vocabulary questions

Our vocabulary data came from 1,417 story readings by 327 of the 364 students described earlier. The number of story readings with vocabulary data averaged 4.33 per student, but ranged from a minimum of 1 to a maximum of 22, with median 3. As we will explain later, each story reading posttested 4 words after having pretested 3 of them.

To disaggregate students by their prior overall vocabulary, we split them into grade levels based on their grade equivalent scores on the Woodcock Reading Mastery Test subtest for Word Comprehension (Woodcock, 1998). This disaggregation is restricted to the 231 students whose test scores were easy to match up with their vocabulary data from the Reading Tutor. Student names and birthdates recorded on test forms often deviated from their enrollment data in the Reading Tutor, necessitating a laborious matching-up process all too familiar to education researchers and data miners. We had used such a match-up in 2002 for the cloze analysis reported in (Mostow *et al.*, 2002c) and in the experiment described previously. Unfortunately, by the time we performed the analyses for the current section as well as the Evaluation of Vocabulary Previews section of this paper, we did not find the results of the original match-up in a form we could reuse with the vocabulary data, and the analyst who had created it was long gone. Part of the problem was that the Reading Tutor recorded data in log files that were unwieldy to parse, aggregate, and analyze. Starting with the fall 2002 version, the Reading Tutor stored data directly to a database (Mostow *et al.*, 2002b) more conducive to such analyses.

Except for one student with 8 trials who scored at a kindergarten level, the students' levels ranged from grade 1 to grade 6, with the number of students at each level ranging from 20 at grade level 1 up to 54 at grade level 3. The mean number of posttested words per student ranged from 8.8 at grade level 1 up to 28.2 at grade level 5.

Item repetition

The data set includes questions on 869 distinct words. The number of students tested on a word ranged from 1 to 39, with median 4 and mean 6.51. Far too few students saw each question to support conventional application of Item Response Theory, which estimates the difficulty of a question based on the

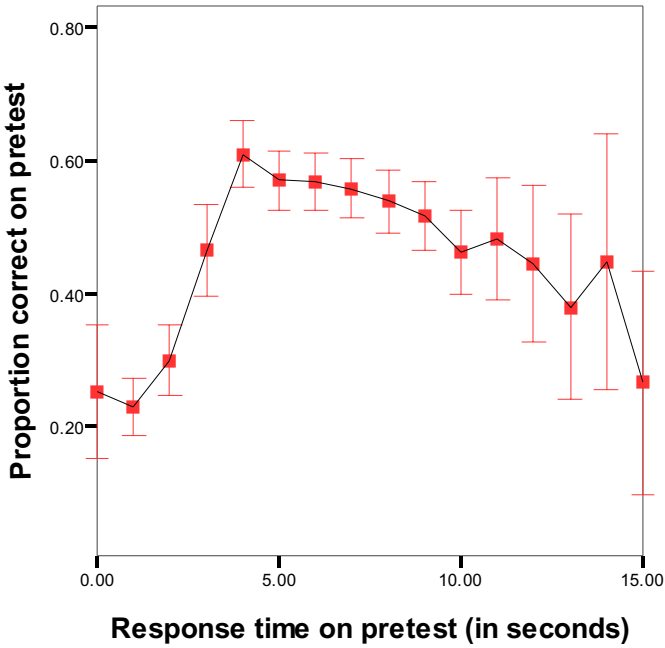

FIGURE 5
Vocabulary pretest performance as a function of response time.

percentage of students who answer it correctly. Only 188 words were tested on 10 or more students. Moreover, the number of students tested on the same word is only an upper bound on the number of students who were asked the same question, because distractors were chosen randomly each time.

Hasty responses

Performance on vocabulary questions varied with response time, just like the within-story cloze questions discussed earlier.

Figure 5 shows the proportion of correct pretest responses as a function of the number of seconds to respond; posttest responses followed a similar pattern. 3.3% of the 4,251 pretest responses took longer than 15 seconds and are not shown. 10.8% of the story pretest responses (204 responses by 92 students) were faster than 2 seconds. So were 16.7% of the 5,668 story posttest responses. These "hasty responses" were at chance level (25%) overall; there were too few per student to tell if individual students' hasty responses were at chance. Accordingly, we treated correct responses faster than 2 seconds as lucky

TABLE 4
Number of thoughtful and hasty pretest and posttest responses

Number of responses	Thoughtful Posttest	Hasty Posttest	Total Posttest
Thoughtful pretest	3372	421	3793
Hasty pretest	204	254	458
Total pretest	3576	675	4251
No pretest	1162	255	1417
Total posttest	4738	930	5668

guesses, and did not credit them as evidence of knowing the word – nor did we treat hasty incorrect responses as evidence of not knowing the word, because 2 seconds was generally too short to answer deliberately.

Table 4 splits hasty posttest responses by whether the pretest response on the same word was hasty. Posttest responses were hasty in 55% of the cases where the pretest response was hasty, but in only 11% of the cases where it was thoughtful.

Reliability

As described earlier in relation to cloze responses, we computed the Guttman split-half reliability of students' pretest performance on the matching task, excluding hasty responses and students with fewer than *m* thoughtful responses, for various values of the threshold *m.* As we increase *m*, the number of students with at least *m* thoughtful responses can only decrease. For $m = 2$, there were 323 students. For $m = 10$, there were 140 students. For $m = 15$, there were 94 students. As *m* increased to 20, the number of students shrank to 62. Corrected split-half reliability increased from .37 for $m = 2$ to .65 for $m = 10$ and .73 for $m = 15$, then down to .69 for $m = 20$ as the set of students became smaller and more homogeneous.

External validity

Was the matching task a valid measure of vocabulary knowledge? To answer this question, we correlated performance on the 4251 pretest questions against grade equivalent Word Comprehension scores on the Woodcock Reading Mastery Test (Woodcock, 1998). We excluded hasty responses, calculated the percentage correct for each student's thoughtful responses, and correlated these percentages against students' Word Comprehension scores,

excluding students with fewer than m thoughtful responses, for some threshold m. The correlation varied with m. For $m = 1$, there were 231 students with WRMT scores, and the correlation was .49. For $m = 10$, there were 113 students and the correlation was .61. As m increased to 20, the set of students shrank to 55, and the correlation decreased to .54, presumably due to a more homogeneously voracious set of readers. Our data set averaged only 17.3 pretest responses per student, with a median of 9, so it is not surprising that they did not predict Word Comprehension scores as well as the cloze-based model described earlier. Nonetheless, they clearly provided a valid, albeit rough, estimate of students' vocabulary.

Construct validity

The multiple-choice vocabulary questions tested the student's ability to match words to short definitions of them. Such a matching task tested receptive rather than productive vocabulary (i.e. recognition, not recall) because it presented both the words and their definitions, in contrast to a task where the student is given one and must generate the other rather than choosing it from a list. To test only word comprehension, not word identification as well, the Reading Tutor read the prompts and choices aloud, in contrast to paper tests (e.g., Woodcock, 1998) where students must identify the words as well as understand them.

A paper version of such a test may present a list of words and a list of definitions, with instructions to match them up. Such a format lets students start with more familiar words, thereby narrowing down the choices as they go, and matching unfamiliar words by process of elimination rather than based on knowledge of their meaning. In contrast, the multiple-choice format controlled the order in which the student matched the words. The student saw only one definition at a time, making it harder to answer later questions by process of elimination, and could not return to earlier questions as on a paper test. Successive questions used the same four word choices, but randomly reordered each time, further hampering the strategy of answering by process of elimination. Thus the interactive multiple-choice format tested knowledge of word meaning better than a paper-format matching task.

The import of a response to a multiple-choice vocabulary question depends on a number of aspects. First, the difficulty of any multiple-choice question depended not only on the familiarity of the choices but also on the similarity among them, and hence the subtlety of the distinctions necessary to distinguish the correct target choice from the distractors. Nagy (1985) identified three levels of difficulty, illustrated in (Aist, 2002) for the target word *traveler*:

1. Choices have different parts of speech, e.g., *eating*, *ancient*, and *happily* (a verb, an adjective, and an adverb).
2. Choices have the same part of speech but different semantic classes, e.g. *antelope*, *mansion*, and *certainty* (an animal, a physical object, and a mental state).
3. Choices have the same semantic class but different core meanings, e.g., *doctor*, *lawyer*, and *president* (all human).

Students acquire knowledge of word meaning incrementally rather than all at once (Beck *et al.*, 2002). Questions at successive levels of difficulty test this knowledge in successively greater depth. The Reading Tutor selected the distractors at random from story words annotated as vocabulary items, so they were likely to be somewhat challenging in terms of prior familiarity. They were generally nouns, verbs, adjectives, or adverbs, rather than articles, prepositions, or other function words. The fact that the words came from the same story increased the chances that they were semantically related in some way, but their syntactic and semantic similarity varied according to the luck of the draw. Consequently a question generated in this way tended to be a hybrid of the three levels rather than a pure level 1, 2, or 3.

The import of a response is asymmetric in the following sense. A correct response did not necessarily imply knowledge of the word. A pure guess had a 25% probability of being correct – higher if the student could eliminate any of the choices based on knowledge of the other words, even with zero knowledge of the correct target word. Moreover, depending on the similarity among the (non-eliminated) choices, even partial knowledge of the word (such as its likely part of speech based on its morphology) might suffice to answer correctly. In contrast, an incorrect response (unless hasty or careless) provides clear evidence that the student did not know or remember the word, at least at whatever depth was tested by the question.

Finally, the very same question – such as *Which word means "duck feathers?" eider-down; bedstead; scarcely; torrents* – might test something quite different, depending on whether it was asked before or after the Reading Tutor taught the word. Prior to such instruction, this question tested the student's knowledge of word meaning by matching the word *eider-down* to its definition. In contrast, once the Reading Tutor taught *eider-down* as meaning *duck feathers*, matching became a paired-associate task that the student might conceivably be able to perform with no understanding of the word, nor of its

definition. The posttest for each word used the same four word choices as the pretest (albeit in rerandomized order). Thus instruction might conceivably teach a student to answer the posttest by associating the prompt with some superficial non-semantic feature of the target word. Potential exposure effects also differed before and after the story. Using story words as distractors ensured recent (though not necessarily equal) exposure to all four words used as choices when the posttest questions were administered. In summary, the significance of a response to a multiple-choice vocabulary question depends on the choices presented, whether the response was correct, and when the question was asked.

EVALUATION OF VOCABULARY PREVIEWS

We had shown that automated questions could be aggregated to assess students' comprehension and vocabulary. But were they sensitive enough to detect differences between the same student's comprehension in different contexts, and changes over time? That is, could they be used to evaluate effects of tutorial actions on students' vocabulary and comprehension?

Previous work on vocabulary assistance and assessment

Comprehending a text requires knowing what the words mean. A reader who can pronounce a word but does not know its meaning, or crucial facts about it, is at a disadvantage in comprehending the text in which it occurs (Stanovich *et al.*, 1992). Children can learn words from contextual cues while reading, but this type of vocabulary development is inefficient because text appropriate for children learning to read does not contain enough challenging vocabulary (Beck & McKeown, 2001). Moreover, contextual cues are often inadequate to infer word meaning, and trying to guess word meaning from context uses cognitive resources that could be applied to comprehension (Beck *et al.*, 2002, p. 43). Finally, students who most need help with vocabulary are least likely to be able to infer meaning from context (Beck *et al.*, 2002, p. 4).

Direct instruction of vocabulary is therefore necessary (NRP, 2000, pp. 4-4), and indeed appears to lead to better reading comprehension (Freebody & Anderson, 1983; NRP, 2000, pp. 4-20). For example, Beck, Perfetti, and McKeown (1982) reported better performance on semantic tasks by 4th graders who received vocabulary instruction than those who did not. As early as kindergarten, explaining unfamiliar words and concepts in context can remediate deficits in vocabulary and background knowledge (Brett *et al.*, 1996; Elley, 1989; Penno *et al.*, 2002).

However, vocabulary learning takes time. Readers learn word meaning over multiple encounters, rather than all at once. Students can understand a word at several levels (Dale, 1965): 1) never encountered the word before; 2) have seen the word but do not know its meaning; 3) can recognize the word's meaning in context; 4) know the word well. Explicit vocabulary instruction can help, but is too time-consuming to cover many words (Beck *et al.*, 2002). We focus here on helping students get the most out of their first encounter of a word, both in terms of understanding it in context and learning it for future use.

Aist (2001b; 2002) investigated such vocabulary acquisition support by presenting short textual "factoids" – comparisons to other words. He embedded a within-subject automated experiment in the 1999 version of the Reading Tutor to compare the effectiveness of reading a factoid just before a new word in a story to simply encountering the word in context without a factoid. The decision whether to explain a word was randomized. The outcome measure was performance on a multiple-choice test on both explained and unexplained words, administered the next day the student used the Reading Tutor. Analysis of over 3,000 randomized trials showed that factoids helped on rare, single-sense words, and suggested that they helped third graders more than second graders (Aist, 2001b). Both the factoids and the multiple-choice questions were generated automatically, with uneven results. Problematic definitions used obscure words students did not know, explained common words they already knew, explained the wrong word sense, or were socially unacceptable (Aist, 2001a, p. 89).

Experimental design

As follow-on work, we embedded a within-subject experiment in the Reading Tutor to evaluate the effectiveness of previewing new vocabulary words before a story. As in (Aist, 2001b), the experiment previewed some new words but not others. Would previewing affect students' subsequent comprehension of those words and of sentences containing them?

Before a student read a story, the Reading Tutor randomly chose from the story four vocabulary words new to the student. "New" meant that the student had not previously encountered these words in the Reading Tutor. "Vocabulary word" meant a "defined" word according to the same criterion used to classify word difficulty for cloze questions, as described earlier — namely, a story word for which a staff member had written a synonym and short definition. The Reading Tutor randomly assigned these four vocabulary words to four different treatments, administered in random order:

- **Control:** Do not pretest or explain word meaning.
- **Test-only:** Pretest but do not explain word meaning.
- **Synonym:** Pretest word meaning, then relate the word to a simpler synonym appropriate to the story.
- **Definition:** Pretest word meaning, then give a short explanation appropriate to the story.

Thus the story reading generated a set of four trials (one trial for each treatment) matched by student and story, with treatment order randomly counterbalanced across trials.

First the Reading Tutor assessed students' self-reported familiarity with all four words. This step was inspired by the "limericks" experiment detailed in (Aist, 2001a, Section 6.2) and summarized in (Aist, 2002). Aist had used questions of the form *Have you ever seen the word "dolorous" before?* to measure students' familiarity with rare words as an experimental outcome on a posttest given one or two days after alternative treatments involving the words. He found that:

- Word familiarity was more sensitive to treatment than word knowledge as tested by matching words to synonyms.
- "Students in lower grades were more likely to (probably incorrectly) report that they had seen a word before that had not been presented in the Reading Tutor: 50% for second grade, 33% for third grade, 29% for fourth grade, and 13% for fifth grade, on the 2 of 8 words they did not see in the study." (Aist, 2001a, pp. 125-126)
- "The relationship between familiarity and word knowledge was stronger and statistically significant for students in the higher grades, but almost zero and not significant in the lower grades." (Aist, 2001a, p. 130)

We rephrased the familiarity question to ask about word meaning, not just prior exposure. The Reading Tutor displayed and read aloud a question of the form *Do you know what SCARCELY means? YES; NO.* It randomized the order of the *YES* and *NO* response choices to control for order bias, and highlighted them in sequence as in the other multiple-choice questions. Besides assessing (self-reported) familiarity with the word, this question served to pronounce the word for the student, in order to scaffold word identification in all four treatment conditions.

To pretest word knowledge, the Reading Tutor used a multiple-choice question of the form described earlier. The prompt gave a short definition, and the choices were the four vocabulary words the Reading Tutor had selected, e.g. *Which word means "barely"? bedstead; eider-down; torrents; scarcely.* To teach a synonym, the Reading Tutor put it in a yes-no question, e.g. *Did you know SCARCELY means BARELY?* To teach a definition, the Reading Tutor displayed and read it aloud to the student, e.g. *TORRENTS means "heavy flows of water."*

After pretesting all words except the control word, and teaching the synonym and definition, the Reading Tutor helped the student read the story, but giving assistance only with decoding, not with word meaning. Now and then (every several minutes, on average) it randomly inserted a cloze question as described earlier. Some of these questions included one or more of the four experimental words, whether as part of the prompt or among the choices, e.g. *There was thunder and lightning, and the rain poured down in _____. torrents; sensitive; gracious; eider-down.* The Reading Tutor then resumed the story by displaying the correct complete sentence for the student to read aloud.

After the story, the Reading Tutor posttested each experimental vocabulary word with a multiple-choice question of the same form as the pretest, as shown in Figure 4. Each question presented the same four words, re-randomizing their order.

Relation of self-reported familiarity to pretest performance

How well did students' self-reported familiarity with word meaning predict their pretest performance on those words? Very poorly. We compared 3,651 thoughtful pretest responses against the responses to *Do you know what WORD means?* The percentage of words that students claimed to understand rose monotonically from 30% at age 7 to 52% at age 12, but their pretest performance on the words they claimed to know fell far short of 100%, rising monotonically from 47% at age 7 to 68% at age 12.

Metacognition includes knowing what you know, and being able to distinguish it from what you don't know. Pretest performance was higher on words students claimed to know than on words they didn't claim to know. The difference between the two percentages reflected metacognitive growth, rising near-monotonically from 7% at age 7 to 20% at age 12.

Effect of preview on vocabulary matching posttest

How did students perform on the posttest questions? We first describe the

overall set of responses; our subsequent analyses take proper account of student identity. The percentage of correct answers on the 5,668 post-story multiple-choice questions was 49% for the Control and Test-only conditions, 53% for Synonym, and 54% for Definition, compared to chance-level performance of 25%. The standard error for these percentages was ±1.3%. Evidently a multiple-choice question pretest on a word without feedback on the student's answer did not matter, even though it exposed the student to the word's definition; posttest performance still averaged 24% higher than chance, the same as for Control words. Explaining a word – whether by definition or by synonym – increased posttest performance to 28% higher than chance. We collapsed the four conditions into two for further analysis:

- **Taught** (Synonym and Definition): explained before the story where students first saw them in the Reading Tutor
- **Untaught** (Control and Test-only): encountered in the story without being explained beforehand

We now analyze how the four types of vocabulary previews affected student performance on the post-story matching task. Did teaching a word improve performance on the posttest?

To answer this question, we constructed a logistic regression model (Menard, 1995) in SPSS to predict performance on each posttest question (correct *vs*. incorrect) as the outcome variable. 930 (16%) of the 5,668 posttest responses were hasty, so we excluded them because hasty responses tested luck rather than knowledge (on average they were at chance level). We included treatment (taught *vs*. untaught) and userID (student identity) as factors, and word frequency as a covariate to control for word difficulty, or at least obscurity.

Including student identity as a factor controls for differences among students, and 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 vocabulary word was taught at most once, and was unlikely to affect the student's knowledge of other vocabulary words.

We neglect possible dependency among responses caused by reusing the same four words (albeit in re-randomized order) as the choices for successive multiple-choice questions. Such dependency might affect students' performance, for example if they had the metacognitive skills to

eliminate some words based on prior or taught knowledge of other words. However, it should neither advantage nor disadvantage any of the treatments, because treatment order was randomized.

Did previews help? Yes. Treatment effects were significant at $p < .001$. So was student identity. Word frequency was not significant ($p = .172$). In short, there were strong main effects both for treatment and for individual differences.

Effect of prior knowledge on vocabulary matching posttest

Next we investigated which words and students benefited from the vocabulary previews. Explaining a word was unlikely to help a student who already understood it. We therefore first disaggregated trials based on prior knowledge of the word, as revealed by the student's response to the pretest question. We excluded the 1,417 control trials because they did not pretest words, and so gave no information about the student's prior knowledge of the word. We excluded the 930 trials with hasty posttest responses as they gave no information about the student's post-story knowledge of the word. This left 3,576 trials:

- **Hasty** (204 trials, 92 students): Responded faster than 2 seconds.
- **Knew** (1,792 trials, 286 students): Answered correctly after 2 or more seconds. This set excludes hasty responses, but includes an unknown percentage of lucky guesses.
- **Didn't know** (1,580 trials, 289 students): Answered incorrectly after 2 or more seconds.

To analyze each subset of trials, we used the same form of logistic regression model as described previously with the pretest data. Treatment was significant ($p < .001$) for the "didn't know" trials, but not for the other two cases. For the "knew" and "didn't know" cases, student identity was significant at $p < .001$. However, for trials with hasty pretest responses, student identity was merely suggestive ($p = .11$), indicating that it doesn't matter much who you are (or what you know) if you respond hastily – or perhaps that hasty pretest responses tended to presage hasty posttest responses as well, as Table 4 showed. Word frequency was not significant for any case ($p > .2$). In sum, previews helped words that students didn't know, at least well enough to match them to their definitions on the pretest.

"Age and ability levels can affect the efficacy of various vocabulary instruction methods" (NRP, 2000, p. 4-27). To identify who benefited from

TABLE 5
Effects on post-story matching task, by student vocabulary level.

Vocabulary grade level according to score in Word Comprehension	% correct if word taught	# trials	# students	% correct if word not taught	# trials	# students	Increase in % correct	Effect size	p for teaching target word
1	32%	44	15	45%	14	10	-13%	-0.32	0.305
2	36%	104	41	30%	55	33	6%	0.16	0.264
3	36%	169	44	35%	103	40	1%	0.03	0.982
4	52%	224	47	41%	123	40	11%	0.34	0.073
5	61%	210	35	42%	96	33	19%	0.51	0.001
6+	72%	80	18	37%	49	18	35%	1.07	0.038

the vocabulary previews, we disaggregated treatment effects by students' grade equivalent scores on the Word Comprehension subtest of the Woodcock Reading Mastery Test (Woodcock, 1998), rounded to the nearest integer. Table 5 shows treatment effects at each grade level for 210 students with known scores, based on 1,271 "didn't know" trials with thoughtful posttest responses. Percentages correct are averaged per-student. Effect size is computed as the difference in percentage correct divided by average within-treatment standard deviation. Significance is taken from the *p* value for treatment in the logistic regression model for students at that grade level for vocabulary.

Treatment effects emerged only at grade level 4 (effect size .34) and became significant only at grade levels 5 (effect size .51) and 6+ (effect size 1.07). In sum, previewing words before a story improved students' ability to match them to their definitions after the story – but only for words they didn't already know, and only if they had at least a grade 4 vocabulary.

Effect of preview on cloze performance during the story

Now we analyze how vocabulary previews affected cloze performance during the story. During the 1,417 story readings with vocabulary previews, the Reading Tutor posed 3,254 "defined word" cloze questions to 201 students. This data set excludes about 15 minutes' worth of data that we discarded for one student because it was corrupted by data duplication.

To test for effects of vocabulary previews on cloze performance, we constructed a logistic regression model. As shown earlier, the outcome for each

cloze item was whether the response was correct. However, we did not exclude hasty cloze responses. An appropriate criterion for "hasty" might need to vary with the length of the cloze question. But the actual reason we included hasty cloze responses is more embarrassing, though (we hope) informative.

The Reading Tutor had logged the data for the vocabulary previews and cloze questions in separate files. Analyzing the effect of previews on cloze performance therefore required something akin to a massive join operation on the two types of files, implemented by developing a special-purpose perl script to put the combined data in a format that SPSS could input. By the time we realized that this script omitted response time for cloze items, the effort to revise and rerun it was not worth the effort. In contrast, the database representation (Mostow *et al.*, 2002b) used by subsequent versions of the Reading Tutor makes it much easier to incorporate such additional features into an analysis after the fact.

As before, we included student identity as a factor to control for individual differences. However, modelling cloze performance was trickier because treatment, prior knowledge, and elapsed time might each affect students' comprehension of the target word, the distractors, or other words in the clozed sentence. Moreover, the pretest provided only partial information about students' prior knowledge of specific words. Finally, we wanted to keep the model as simple as possible to avoid over-fitting the data. This constraint precluded explicitly modelling every possible combination of treatment and prior knowledge for the target word, distractors, and sentence words.

Accordingly, we summarized students' prior knowledge of the target word and distractors by counting how many of them they got *wrong* on the pretest (not counting hasty responses), because incorrect responses demonstrated ignorance more unambiguously than correct responses proved knowledge, as discussed previously. We likewise counted the number of other words in the clozed sentence that students got wrong on the pretest. Similarly, we summarized treatment as the number of response choices and sentence words, respectively, taught prior to the story, whether by definition or by synonym. Thus the model included as covariates the number of unknown choices, the number of unknown sentence words, the number of taught choices, and the number of taught sentence words.

To illustrate, consider the next to last cloze question in Table 1: *And the very next day, the _____ had turned into a lovely flower. grain; lily; walnut; prepare.* Suppose that this cloze question occurred during a story reading for which the Reading Tutor had assigned *dote*, *lovely*, *grain,* and *walnut* to the Control, Test-Only, Synonym, and Definition treatments described earlier. Also suppose that the student's pretest responses before the story were

incorrect for *lovely* and *grain* but correct for *walnut*. Then the number of unknown choices would be one (*grain*), the number of unknown sentence words would be one (*lovely*), the number of taught choices would be two (*grain* and *walnut*), and the number of taught sentence words would be zero.

We constructed a table of cloze performance contingent on these four variables. The data were distributed unevenly. Consequently, the percentage of correct responses was estimated much better in some cells than in others. The bulk of the target words for the cloze questions were neither pretested nor taught, because they were not among the four vocabulary words selected from the story for the matched trials. The number of unknown or taught sentence words was mostly zero. The number of unknown or taught choice words was spread more evenly. Other things being equal, performance typically (but not always) dropped a few percent for each unknown choice word and rose a few percent for each choice word taught.

Unlike the vocabulary posttest questions at the end of the story, cloze questions could appear at any time during the story. They might therefore be affected differentially by memory decay effects. Accordingly, our logistic regression model also included as a covariate the time (in minutes) from the start of the story until the cloze question was asked.

Table 6 shows the results of the logistic regression. The next to last column shows the statistical significance of each predictor variable. The last column measures the influence of each covariate. The Beta value for a covariate shows how an increase of 1 in the value of the covariate affected the log odds of the outcome. Thus each additional unknown response choice reduced the log odds of a correct response by 0.228. This prior knowledge effect was significant at $p < .001$. Conversely, each additional taught choice increased the log odds of a correct response by 0.154. Prior knowledge and treatment affected sentence words like response choices, albeit not significantly. The effect of elapsed time was slightly positive, opposite of the expected decay, but was not statistically significant. Student identity was a highly significant predictor of correct response ($p < .001$).

In short, knowing or being taught a new word before a story increased the student's chances of correctly answering a cloze question that involved the word. These knowledge and treatment effects were significant if the word was one of the choices (target word or distractor), and were insignificant but positive trends for other words in the cloze sentence. The model tests for knowledge and treatment effects, but not for interactions between them, as a more sensitive model might do.

TABLE 6
Effects of pretest knowledge and treatment on cloze performance

Feature	Chi-square	DF	p value	Beta
# unknown response choices	11.2825	1	0.0008	-0.228
# unknown sentence words	1.49969	1	0.2207	-0.340
# taught response choices	6.12119	1	0.0134	0.154
# taught sentence words	2.15670	1	0.1419	0.346
cloze time - story start time	1.67847	1	0.1951	0.013
student identity	420.650	201	0.0000	

Table 7 shows how cloze performance, averaged per-student, varied by grade level and whether any choices were taught. Effect sizes are computed as difference in mean percentage correct, divided by average within-condition standard deviation. Significance levels are *p* values from a similar logistic regression model for students with a given vocabulary grade level, but with the number of cloze choices taught encoded as a binary feature (zero vs. 1 or more) instead of as a covariate, so as to correspond more closely to the comparisons in the table.

Treatment effects of vocabulary previews on cloze performance for students with Word Comprehension scores in grades 1, 2, or 3 were positive, with similar effect sizes, so we aggregated them together to see if the overall effect was significant. Treatment for the grade 1-3 range overall was statistically significant at $p = .022$. In contrast, treatment effects at grade levels 4, 5, and 6+ were small or negative, and statistically insignificant ($p > .5$).

CONCLUSION

We have described and evaluated the automated generation of questions to assess students' vocabulary and comprehension in Project LISTEN's Reading Tutor. A model using the automatically generated multiple-choice cloze questions predicted Word Identification, Word Comprehension, and Passage Comprehension scores on the Woodcock Reading Mastery Test (Woodcock, 1998) with high reliability and correlation exceeding .8, even for students the model was not tuned on. For the subset of students who answered 10 or more items, a multiple-choice format for matching words to their definitions achieved reliability .65 and external validity of .61 for Word Comprehension

TABLE 7
Cloze performance by treatment and grade level

Vocabulary grade level according to score in Word Comprehension	% correct if ≥ 1 choices taught	# trial	# students	% correct if no choices taught	# trials	# students	Increase in % correct	Effect size	p for teaching choices
1	36%	67	13	25%	21	8	11%	0.31	0.381
2	33%	176	34	27%	80	20	7%	0.21	0.094
3	37%	385	47	32%	147	40	5%	0.19	0.178
levels 1-3	35%	628	94	29%	248	68	6%	0.21	0.022
4	45%	464	46	43%	317	40	2%	0.08	0.998
5	49%	403	33	54%	203	28	-5%	-0.21	0.561
6+	54%	252	22	57%	160	20	-2%	-0.09	0.896
levels 4-6+	48%	1119	101	50%	680	88	-1%	-0.05	0.64

scores. It is possible to predict comprehension just as well from other data, such as oral reading fluency (Deno, 1985) or speech recognizer output and student help requests (Beck *et al.*, 2003, 2004; Jia *et al.*, 2002). However, the automated questions offer arguably better construct validity for specifically measuring comprehension as opposed to more general reading processes.

We used both matching and cloze questions to analyze the benefits of brief textual explanations of new vocabulary. Vocabulary previews before a story enabled students to match words to definitions after the story significantly better than just encountering the words in the story — but only for students at or above a grade 4 vocabulary level, and only for words they missed on the pretest. Effect sizes rose from .34 at grade 4 to 1.07 at grade 6. Such previews boosted performance during the story on cloze questions involving the word, whether as cloze target or distractor – but only for students at a grade 1-3 vocabulary level. Assistive effects of treatment on cloze performance at these levels were statistically significant, with effect sizes of .2 to .3.

Why did the two types of automated questions give such opposite results as to which students benefited from previews? It seems unlikely that vocabulary previews would help children at a grade 1-3 level integrate vocabulary and context to answer cloze questions, yet not improve their performance on the presumably easier paired-associate task of matching words to their definitions.

The probability that the cloze result for students with a grade 1-3 vocabulary was a statistical fluke is $p = 0.022$ – higher if corrected for any multiple comparisons entailed in aggregating grades 1-3. Or possibly this result is an artifact of some statistical bias that we failed to cancel out even by controlling for student identity as a factor in the logistic regression model. If the grade 1-3 cloze result was illusory, we can simply interpret our body of results as evidence that the vocabulary previews did not help below grade 4, and had no measurable effect on cloze performance.

But what if the grade 1-3 cloze result was real? How then to explain the negative results both for grade 4-6 cloze and for grade 1-3 matching? The non-impact on cloze above grade 4 might be due to a ceiling effect in which students already did as well as they could, albeit not perfectly. For example, perhaps they already knew enough of the previewed words to answer the subset of cloze questions where previews might otherwise have helped them.

However, the impact in grade 1-3 of previews on cloze – but not on matching – is harder to explain. What if we are wrong about the relative difficulty of the cloze and matching tasks, due to an "expert blind spot" like the one that leads math teachers to mis-predict the relative difficulty of equations and word problems for young students (Nathan & Koedinger, 2000)? Could it actually be easier for children at this level to use vocabulary knowledge to answer cloze questions than to match words to decontextualized definitions? Further work is required to replicate and explain this intriguing discrepancy.

Acknowledgements

This work was supported by the National Science Foundation under Grant Numbers REC-9979894 and REC-0326153. 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. Portions of this work were first reported at a workshop (Mostow et al., 2002c) and a conference (Mostow et al., 2003b). We thank other members of Project LISTEN who contributed to this work, the students and educators at the schools where Reading Tutors recorded data, Larry Hedges for statistical expertise, and Ryan Baker and Val Shute for comments on an earlier draft. We alone are responsible for any errors or shortcomings.

REFERENCES (See also: www.cs.cmu.edu/~listen)

Aist, G. (2000, August). Identifying words to explain to a reader: A preliminary study. *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, Austin, TX, 1061.

—-. (2001a). *Helping Children Learn Vocabulary during Computer-Assisted Oral Reading.* Unpublished Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA.

—-. (2001b). Towards automatic glossarization: Automatically constructing and administering vocabulary assistance factoids and multiple-choice assessment. *International Journal of Artificial Intelligence in Education, 12*, 212-231.

—-. (2002). Helping Children Learn Vocabulary During Computer-Assisted Oral Reading. *Educational Technology and Society, 5*(2), http://ifets.ieee.org/periodical/vol_2_2002/aist.html.

Aist, G., & Mostow, J. (2004, in press). Faster, better task choice in a reading tutor that listens. In V. M. Holland & F. N. Fisher (Eds.), *Speech Technology for Language Learning*. Lisse, The Netherlands: Swets & Zeitlinger Publishers.

Beck, I. L., & McKeown, M. G. (2001). Text Talk: Capturing the Benefits of Read-Aloud Experiences for Young Children. *The Reading Teacher, 55*, 10-20.

Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing Words to Life: Robust Vocabulary Instruction*. NY: Guilford.

Beck, I. L., Perfetti, C. A., & McKeown, M. G. (1982). Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology, 74*(4), 506-521.

Beck, J. E., Jia, P., & Mostow, J. (2003, June 22-26). Assessing student proficiency in a Reading Tutor that listens. *Proceedings of the 9th International Conference on User Modeling*, Johnstown, PA, 323-327.

—-. (2004). Automatically assessing oral reading fluency in a computer tutor that listens. *Technology, Instruction, Cognition and Learning,* 2, 61-81.

Brett, A., Rothlein, L., & Hurley, M. E. (1996). Vocabulary Acquisition from Listening to Stories and Explanations of Target Words. *Elementary School Journal, 96*(4), 415-422.

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.

Dale, E. (1965). Vocabulary measurement: Techniques and major findings. *Elementary English, 42*, 82-88.

Deno, S. L. (1985). Curriculum-Based Measurement: The emerging alternative. *Exceptional Children, 52*(3), 219-232.

Dolch, E. (1936). A basic sight vocabulary. *Elementary School Journal, 36*, 456-460.

Drott, M. C. (n.d.). *The Cloze Test (free software)*. Retrieved April 3, 2002, from http://drott.cis.drexel.edu/clozeproze.htm

Elley, W. B. (1989). Vocabulary acquisition from listening to stories. *Reading Research Quarterly, 24*, 174-187.

Entin, E. B. (1984). Using the cloze procedure to assess program reading comprehension. *SIGCSE Bulletin, 16*(1), 448.

Freebody, P., & Anderson, R. C. (1983). Effects of vocabulary difficulty, text cohesion, and schema availability on reading comprehension. *Reading Research Quarterly, 18*(3), 277-294.

Jia, P., Beck, J. E., & Mostow, J. (2002, June 3). Can a Reading Tutor that Listens use Inter-word Latency to Assess a Student's Reading Ability? *Proceedings of the ITS 2002 Workshop on Creating Valid Diagnostic Assessments*, San Sebastian, Spain, 23-32.

Johns, J. L. (1977, December 1-3). An Investigation Using the Cloze Procedure as a Teaching Technique. *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.

McKeown, M. G. (1993). Creating effective definitions for young word learners. *Reading Research Quarterly, 28*(1), 17-31.

Menard, S. (1995). Applied Logistic Regression Analysis. *Quantitative Applications in the Social Sciences, 106*.

Mostow, J., & Aist, G. (2001). Evaluating tutors that listen: An overview of Project LISTEN. In K. Forbus & P. Feltovich (Eds.), *Smart Machines in Education* (pp. 169-234). Menlo Park, CA: MIT/AAAI Press.

Mostow, J., Aist, G., Bey, J., Burkhead, P., Cuneo, A., Junker, B., Rossbach, S., Tobin, B., Valeri, J., & Wilson, S. (2002a, June 27-30). Independent practice versus computer-guided oral reading: Equal-time comparison of sustained silent reading to an automated reading tutor that listens. *Ninth Annual Meeting of the Society for the Scientific Study of Reading*, Chicago, Illinois.

Mostow, J., Aist, G., Burkhead, P., Corbett, A., Cuneo, A., Eitelman, S., Huang, C., Junker, B., Sklar, M. B., & Tobin, B. (2003a). Evaluation of an automated Reading Tutor that listens: Comparison to human tutoring and classroom instruction. *Journal of Educational Computing Research, 29*(1), 61-117.

Mostow, J., Aist, G., Huang, C., Junker, B., Kennedy, R., Lan, H., Latimer, D., O'Connor, R., Tassone, R., Tobin, B., & Wierman, A. (2004, in press). 4-Month Evaluation of a Learner-controlled Reading Tutor that Listens. In V. M. Holland & F. N. Fisher (Eds.), *Speech Technology for Language Learning*. Lisse, The Netherlands: Swets & Zeitlinger Publishers.

Mostow, J., Beck, J., Bey, J., Cuneo, A., Sison, J., & Tobin, B. (2003b, June 12-15). An Embedded Experiment to Evaluate the Effectiveness of Vocabulary Previews in an Automated Reading Tutor. *Tenth Annual Meeting of the Society for Scientific Studies of Reading*, Boulder, CO.

Mostow, J., Beck, J., Chalasani, R., Cuneo, A., & Jia, P. (2002b, October 14-16). Viewing and Analyzing Multimodal Human-computer Tutorial Dialogue: A Database Approach. *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces (ICMI 2002)*, Pittsburgh, PA, 129-134.

Mostow, J., Tobin, B., & Cuneo, A. (2002c, June 3). Automated comprehension assessment in a reading tutor. *Proceedings of the ITS 2002 Workshop on Creating Valid Diagnostic Assessments*, San Sebastian, Spain, 52-63.

Nagy, W., Herman, P., & Anderson, R. (1985). Learning words from context. *Reading Research Quarterly, 20*, 233-253.

Nathan, M. J., & Koedinger, K. R. (2000). An investigation of teachers' beliefs of students'algebra development. *Cognition and Instruction, 18*(2), 207-237.

NRP. (2000). *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* (No. 00-4769). Washington, DC: National Institute of Child Health & Human Development.

Penno, J. F., Wilkinson, I. A. G., & Moore, D. W. (2002). Vocabulary Acquisition from Teacher Explanation and Repeated Listening to Stories: Do They Overcome the Matthew Effect? *Journal of Educational Psychology, 94*(1), 23-33.

Spache, G. D. (1981). *Diagnostic Reading Scales*. Monterey, CA: McGraw-Hill.

SPSS. (2000). SPSS for Windows (Version 10.1.0). Chicago, IL: SPSS Inc.

Stanovich, K., West, R., & Cunningham, A. E. (1992). Beyond phonological processes: Print exposure and orthographic processing. In S. Brady & D. Shankweiler (Eds.), *Phonological Processes in Literacy*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Vacca, J. A. L., Vacca, R. T., & Gove, M. K. (1991). *Reading and Learning to Read* (2nd ed.). New York: Harper Collins.

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.