Helping Children Learn Vocabulary during Computer-Assisted Oral Reading

Gregory Aist

December 12, 2000
CMU-LTI-00-167

Language Technologies Institute, School of Computer Science,
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213-3720

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in Language and Information Technologies

Committee:
Jack Mostow, mostow@cs.cmu.edu, Robotics Institute, Language Technologies Institute, Human-Computer Interaction Institute, and Center for Automated Learning and Discovery, advisor
Albert Corbett, al.corbett@cs.cmu.edu, Human-Computer Interaction Institute
Alex Rudnicky, air@cs.cmu.edu, Computer Science Department and Language Technologies Institute
Charles Perfetti, perfetti+@pitt.edu, Psychology Department, Linguistics Department,
and Learning Research and Development Center (LRDC), University of Pittsburgh

Abstract

This dissertation addresses an indispensable skill using a unique method to teach a critical component: helping children learn to read by using computer-assisted oral reading to help children learn vocabulary. We build on Project LISTEN's Reading Tutor, a computer program that adapts automatic speech recognition to listen to children read aloud, and helps them learn to read (http://www.cs.cmu.edu/~listen). To learn a word from reading with the Reading Tutor, students must encounter the word and learn the meaning of the word in context. We modified the Reading Tutor first to help students encounter new words and then to help them learn the meanings of new words. We then compared the Reading Tutor to classroom instruction and to human-assisted oral reading. The result: Second graders did about the same on word comprehension in all three conditions. However, third graders who read with the 1999 Reading Tutor, modified as described in this dissertation, performed better than other third graders in a classroom control on word comprehension gains – and even comparably with other third graders who read one-on-one with human tutors.

Story choice. In the spring of 1998, 24 students in grades 2, 4, and 5 at a low-income urban elementary school used the Reading Tutor with a student-only story choice policy. In the fall of 1999, 60 students in grades 2 and 3 at a (different) low- to middle-income urban elementary school used a revised version in which the Reading Tutor and the student took turns picking stories. The students who used the Take Turns Reading Tutor in fall 1999 averaged 64.1% new sentences out of ~35,000 sentences overall, calculated on a per-student basis. This was a significantly higher percentage of new material than the 60.1% for the ~10,000 sentences read by the students who used the student-only story choice policy Reading Tutor in spring 1998. Furthermore, the Reading Tutor's story choices helped the most for those who did not choose new stories themselves: about half of the students picked new stories less than half the time on their own turns, with some choosing as few as 15% new stories. With the Reading Tutor's choices included, all students read about 50% or more new stories.

Vocabulary help. By augmenting stories with vocabulary help such as short context-specific explanations or comparisons to other words, the Reading Tutor can help students learn words better than they would from simply reading the unaugmented stories.

We augmented text with "factoids": automatically constructed comparisons of a target word to a different word drawn from WordNet, an electronic lexical database. A four-month study conducted in Fall 1999 compared text with vs. text without factoids. A control trial consisted of a student seeing a target word in a sentence and – on a later day – answering an automatically constructed multiple choice vocabulary question on the target word. An experimental trial inserted a factoid prior to presenting the sentence containing the target word. In total, over 3000 trials were completed. There was no significant difference overall between experimental and control conditions; however, exploratory analysis identified conditions in which factoids might help. In particular, story plus factoid was more effective than story alone for the 189 trials on single-sense, rare words tested one or two days later (44.1% ± s.e. 37.7% vs. 25.8% ± s.e. 29.4%, $p < .05$ prior to correction for multiple comparisons). Story plus factoid was also more effective than story alone for third graders seeing rare words (42.0% ± s.e. 28.4% vs. 36.2% ± s.e. 22.9%, $p < .10$ prior to correction). The suspected benefit of seeing the factoid was perhaps due to a word recency effect – sometimes the comparison word was the correct answer in the multiple choice question.

Comparison to classroom instruction and human-assisted oral reading. Human tutors are often considered the gold standard for instruction, and while computer instruction can (sometimes) beat classroom instruction, it typically falls well short of one-on-one human tutoring. In a year-long study, 144 second and third graders at an urban elementary school received classroom instruction for most of the school day, along with one of three 20-minute-per-day treatments. Students were assigned to exactly one of: (a) reading (and writing) with the Reading Tutor, (b) reading (and writing) with a human tutor, or (c) continuing with regular classroom instruction. All three treatment conditions included a range of activities, including some directed at vocabulary development. Thus we were comparing three comprehensive treatments on a single aspect of learning to read, not three treatments aimed specifically at encouraging vocabulary development. Students were pre-tested and post-tested on the Woodcock Reading Mastery Test, a norm-referenced, professionally administered reading test with subtests measuring Word Attack, Word Identification, Word Comprehension, and Passage Comprehension. Students were also tested on oral reading fluency. This dissertation focuses on vocabulary learning, so we only report results on Word Comprehension. For second graders, all three conditions were comparable. For third graders, results were as follows. The 1999 Reading Tutor, with Take Turns and factoids, achieved an effect size of 0.56 over classroom instruction on Word Comprehension gains ($p = .042$). Human tutors achieved an effect size of 0.72 over classroom instruction ($p = .039$). There was no significant difference between human tutors and the Reading Tutor on Word Comprehension gains.

Follow-on experiments explored ways to make vocabulary assistance even more effective, such as adding short child-friendly explanations to text. An initial test confirmed that even low-reading students could understand short explanations well enough to do better on immediate multiple-choice questions than without such explanations. A within-subject experiment in summer 2000 measured word familiarity and word knowledge on eight (difficult) words with a paper test given one or two days after exposure to those words in one of four conditions: no exposure, definition alone, children's limerick alone, or definition plus children's limerick. Definitions increased all students' familiarity with the words, and limericks yielded a strong trend favoring increased familiarity. Also, while 2nd and 3rd graders performed essentially at chance on word knowledge, 4th and 5th graders learned enough from reading stories and definitions with the Reading Tutor to do better on word knowledge. This study furthermore ruled out the word recency effect as an explanation, since none of the words in the definitions or limerick showed up as answers on the multiple choice test. This experiment also shed light on the relationship between word familiarity and word knowledge: the correlation between word familiarity and knowledge was larger in higher grades. Limericks may have been more effective at strengthening the tie between word familiarity and word knowledge – a direction for future research.

Summary. First, taking turns picking stories helped children see more new sentences and more new stories than they would on their own. Second, augmenting stories with automatically constructed vocabulary assistance helped children learn more from their initial encounters with words than just the story alone – at least, for single-sense rare words tested one or two days later. Follow-on experiments point the way to even better vocabulary assistance. Finally, at least for third graders, the 1999-2000 Reading Tutor with Take Turns and factoids outperformed a classroom control on Word Comprehension gains – and was even competitive with one-on-one human-assisted oral reading.

This dissertation is online as PDF and Word files at http://www.cs.cmu.edu/~aist/Aist-PhD-dissertation.html

Table of Contents

Acknowledgements

I would like to first acknowledge my committee: Jack Mostow (advisor), Albert Corbett, Alex Rudnicky, and Charles Perfetti. Project LISTEN team members also helped provide an excellent platform for doing this research, including Jessica Abroms, Daniel Barritt, Juliet Bey, Paul Burkhead, Peggy Chan, Andrew Cuneo, Laura Dabbish, Susan Eitelman, James Fogarty, Sreekar Gadde, Rachel Gockley, Mary Hart, Jeff Hill, Cathy Huang, Tzee-Ming Huang, Kerry Ishizaki, Rebecca Kennedy, Andrew Kim, John Kominek, Hua Lan, DeWitt Latimer IV, David Matsumoto, Joshua McConnell, Jennifer Marie Matvya, Chas Murray, Brian Nagy, Sharon Pegher, Cheryl Platz, Amanda Pyles, Susan Rossbach, Mary Beth Sklar, David Steck, Yinglan Tan, Regina Tassone, Brian Tobin, Joe Valeri, Adam Wierman, Sara Wilson, and Calvin Yeung. The predecessor to the Reading Tutor – the Reading Coach – was the fruit of years of work by a previous team; while I did not overlap with them, I built on the foundation they laid. The Reading Tutor uses text from many sources, including text from *Weekly Reader* (a newspaper for children), used with permission. The Reading Tutor adapts Carnegie Mellon's Sphinx II speech recognizer – a result of years of effort by many people in the Carnegie Mellon Speech Group. Dr. Rollanda O'Connor (University of Pittsburgh) has assisted Project LISTEN on various questions, including choice of tests for measuring reading outcomes. Elementary school teacher Fran Ferrara also provided advice. Richard Olson, Helen Datta, and Barbara Wise provided helpful feedback during a visit to the University of Colorado at Boulder in 1997. During my time at Carnegie Mellon University I have benefited from discussions with many Carnegie Mellon University faculty members, including Jaime Carbonell, Bob Frederking, Brian Junker, Rob Kass, Ken Koedinger, Jay McClelland, Raj Reddy, and Larry Wasserman, and also with University of Pittsburgh faculty members including Kurt van Lehn and Alan Lesgold. Friends

and colleagues have provided support, encouragement, and advice, including my classmates Rosie Jones and Klaus Zechner; fellow graduate students Kathy Baker, Suresh Bhavnani, Neil Heffernan, Marsal Gavalda, Jade Goldstein, Santosh Mathan, Daniel Morris, Kerry Ojakian, Paul Placeway, Klaus Ries, Laura Tomokiyo, and Yan Qu. During a visit to Microsoft Research (Redmond, Washington) in 1998 I worked with colleagues at CMU and Microsoft on training acoustic models on children's speech. During an extended visit to Macquarie University (Sydney, Australia) in 1998 I was privileged to interact with many people on my thesis as well as on other topics, especially my host Sandra Williams; also Robert Dale, Marc Dras, and Steve Green. My undergraduate senior honors advisor, Gene Chase, encouraged me towards research in natural language processing. Finally, I would like to thank my family, friends, and loved ones for all the support and encouragement they have given me over the years. You are written on the pages of my heart.

This material is based upon work supported in part by the National Science Foundation under Grants IRI-9505156, CDA-9616546, REC-9720348 (Jay McClelland, P.I.) and IERI-9979894, and by the author's National Science Foundation Graduate Fellowship and Harvey Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) 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.

Project LISTEN's home page is at http://www.cs.cmu.edu/~listen. My curriculum vita may be found at http://www.cs.cmu.edu/~aist/cv.html.

1 Introduction

This dissertation addresses an indispensable skill by using a unique method to teach a critical component: helping children learn to read by using computer-assisted oral reading to help children learn vocabulary. Why should you read this dissertation? Literacy matters: The increasing demands of the information economy require higher and higher standards of reading ability from *everyone*, not just the privileged few. There is a crying need for better tools for literacy development: The United States Department of Education's National Assessment of Educational Progress reported that 69% of American fourth graders read below desired proficiency; 38% were below even the basic level (Donahue et al., 1999). Vocabulary knowledge plays a critical role in reading, by enabling and facilitating comprehension (Snow, Burns, and Griffin 1998). Using computers to boost vocabulary learning holds promise for offering children frequent, engaging encounters with the meanings of words.

We select a particular area of vocabulary learning as follows. First, we focus on learning words during assisted oral reading. Second, we concentrate on initial encounters with words. Third, we subdivide vocabulary learning from initial encounters in text into two stages: encountering new words in text, and learning from those encounters. We demonstrate improvements over baseline computer-assisted oral reading, by: (a) making sure that all students – not just better students – see new text; and by (b) adding information to text so that children can learn more from encounters with words than they would from the original text alone.

Our novel approach builds on a recent advance in computer technology as applied to reading: computer-assisted oral reading. We leverage others' work by building on a software platform representing years of multidisciplinary endeavor: Project LISTEN's Reading Tutor, a computer tutor that listens to children read aloud, and helps them learn to read (Mostow & Aist FF 2001).

We situate our work in real classrooms at two Pittsburgh-area schools: Fort Pitt Elementary School, in a low-income neighborhood of Pittsburgh, and Centennial Elementary School, in a moderate-to-low income urban neighborhood near Pittsburgh. Computers' book-keeping capability enables us to carry out finely detailed in-classroom experiments with massive samples recorded in excruciating detail.

Our results apply to several fields of research. For those interested in computer-assisted oral reading, we demonstrate improvements over Project LISTEN's baseline system prior to our dissertation research. For those working on intelligent tutoring systems, we operationalize a hybrid method for deciding which task to work on next: taking turns. For reading researchers, our experiments illuminate the relative merits of natural text and artificially constructed vocabulary help, and provide an example of automatically generated assessment.

We begin at the beginning: what children need in order to learn to read.

1.1 Learning to read

Reading runs deeper than merely turning print into sound; reading makes meaning from print. What does it take for children to learn to read?

Motivation. Motivation may affect the time a child spends reading. Seeing parents, an older sibling, or classmates read may inspire children to want to read. Having choice in what to read may encourage children to read more. Embarrassment in front of peers or frustration may decrease motivation to read.

Opportunity. Some children have parents who read to them, communities that support public libraries, and a cornucopia of books to read at home. Others have scarcely seen a book prior to kindergarten. In fact, lower exposure to print can be a sad reality for low socioeconomic students even in the classroom (Duke 2000).

Skills. Children must acquire a wide range of skills to ultimately comprehend text (NRP 2000, Snow et al. 1998). *Phonemic awareness* allows children to distinguish and manipulate individual sounds in spoken words. Knowledge of *print conventions* enables children to work with text as placed on a page – for English, left-to-right, top-to-bottom. Mastery of the *alphabetic principle* reveals that individual sounds are written with letters or letter patterns. *Decoding skills* codify how to turn printed letters into sounds. Increased *fluency* leads to faster and more automatic reading. *Background knowledge* increases text understanding. *Vocabulary knowledge* is critical for comprehension. *Drawing inferences* from text and *integrating information* from multiple sources finally allow the reader to make meaning from print.

We focus in this dissertation on learning vocabulary.

1.2 Knowing a word

What does it mean to know a word? A person's knowledge of a word may range from none at all to complete mastery. Aspects of word knowledge include:

Pronunciation. *astronaut* is pronounced [æ s t ɹ o n ɔ t], as written in the International Phonetic Alphabet.

Spelling. *astronaut* is spelled A S T R O N A U T.

Part of speech. *astronaut* is a noun.

Morphology. Inflectional morphology carries agreement. For example, the plural of *astronaut* is *astronauts*. Derivational morphology turns one word into another. For example, *astronaut* (noun) + *-ic* → *astronautic* (adjective).

Syntax. *astronaut* refers to a person, so the word *astronaut* takes *he* or *she* as a pronoun.

Lexical semantics. The core meaning of *astronaut* is a space traveler.

Pragmatics. *cosmonaut* shares the core meaning of *astronaut*, but is used for Russian astronauts.

In this dissertation, we focus on learning the core meanings of words.

1.3 Learning the meaning of a new word

How can we help children learn new words? We consider two primary methods: direct instruction and learning through reading; and a hybrid: adding information to text.

Direct instruction. Intensive study of specific vocabulary words results in solid knowledge of the taught words, but at a high cost in time. For example, a 1983 study taught fourth graders 104 words over a five-month period, with 75 lessons of approximately 30 minutes each – on average about 21 minutes of instructional time per target word (McKeown et al. 1983). Exposures were during varied tasks: "matching words and definitions, associating a word with a context, creating contexts for words, and comparing and contrasting words to discover relationships" (McKeown et al. 1983). In the high-exposure group of words, students saw 26-40 exposures; even for the low-exposure words, students saw 10-18 exposures – a substantial amount of instructional time. Beck and McKeown (1991) suggest that "the problem that effective instruction takes time can be alleviated by targeting instruction toward the most useful words" (Beck and McKeown 1991). Which words? They suggest second-tier vocabulary (McKeown 1993), that is, words that are "of high frequency in a mature vocabulary and of broad utility across domains of knowledge" (Beck and McKeown 1991). Thus, direct instruction may play a role for certain critical words (Zechmeister et al. 1995). Nonetheless, a full-fledged instructional lesson is too time-consuming to use for every new word.

Reading. Children can learn words from written contexts (Nagy, Herman, and Anderson 1985, McKeown 1985, Gipe and Arnold 1978), but the process is incremental. That is, the amount learned from each exposure may be small, but the net effect is still substantial (Eller, Pappas, and

Brown 1988). Also, readers with better vocabularies learn more from context – because of broader and deeper prior knowledge of words – even though less of the information in the text is new to them than to readers with poorer vocabularies (Shefelbine 1990).

Reading offers hope for spurring vocabulary growth – if children can be guided to read material that does in fact contain unknown words. Carver (1994) argues that “students must read books above their independent level in order to consistently run into unknown words, that is, about 1, 2, or 3 unknown words for each 100 words of text”. Easier text simply does not contain enough new words to substantially improve children’s vocabulary (Carver 1994).

Is simple exposure to text sufficient for all readers to learn new words? Perhaps – or perhaps not. McKeown (1985) studied how high- and low-ability students learn words from context. McKeown’s (1985) study examined 15 fifth-graders who, at the end of fourth grade, had scored between grade equivalent 3.3 and grade equivalent 4.1 on the Vocabulary section of the Stanford Achievement Test (Madden et al. 1973). These low-reading fifth graders had trouble learning words from context partly because of incorrect inferences about the meaning of a word from context. One might expect that multiple sentence contexts would eliminate incorrect inferences – not the case. Both the low-reading fifth graders *and* the 15 higher-ability students in McKeown’s (1985) study, who had scored above grade equivalent 4.8 on the Stanford Vocabulary subtest, had some trouble integrating multiple sentence contexts to derive meaning.

There has been some work aimed at teaching children how to learn words from context, but the major effect may be due to practice at learning new words from context and not due to teaching a specific strategy (Kuhn and Stahl 1998). Kuhn and Stahl conclude that “Ultimately, increasing the amount of reading that children do seems to be the most reliable approach to improving their knowledge of word meanings, with or without additional training in learning words from context.” As Schwanenflugel, Stahl, and McFalls (1997) put it, “… the vast majority of a person’s word growth can be accounted for by exposure to words in written and oral contexts, not through direct instruction of some sort, but individual encounters with a word in a natural context are not likely to yield much useful information about that word.”

Adding information to text. Can the context in which a word appears be augmented in some way to make it more useful for learning the word? Typical dictionary definitions may not be written to suit the learner’s needs; explanations written to convey the core sense of the word in plain language work better (McKeown 1993). Presenting context-specific definitions in computer-mediated text has been shown to be helpful for vocabulary acquisition, at least for sixth graders (Reinking and Rickman 1990.) Adding information to text is a hybrid of direct instruction and learning from reading text: first, start with a text to read; second, add brief, targeted instruction about words to the text.

How can we test whether a particular kind of vocabulary assistance helps? We now discuss how to assess vocabulary.

1.4 Assessing vocabulary knowledge

Assessing vocabulary knowledge is a difficult problem because we must sample along two dimensions: first, we must select a subset of words to test out of all the words in English; second, we must select a subset of aspects of word knowledge to test. For example, we could select 20 words at random from a dictionary, and then choose to test children’s ability to define those words or use them in a sentence. In fact, the National Reading Panel recommended that experimenters employ custom-made tests to assess vocabulary – rather than relying solely on standardized tests (NRP 2000).

In this dissertation, we use a variety of methods for selecting words to assess, and for assessing children's knowledge of those words. In Chapter 4, we focus on words with few senses – the "low-hanging fruit" for use in automatically generated vocabulary help and multiple-choice questions. In Chapter 5, we use a well-known published measure of vocabulary knowledge, the Word Comprehension subsection of the Woodcock Reading Mastery Test (American Guidance Service, n.d.; see Chapter 5 for details). In Chapter 6, we first focus on domain-specific content words, testing with an experimenter-written matching task; we then look at domain-independent (but very rare) words, testing with experimenter-written multiple choice questions.

Now we have discussed what it means to know a word, and how to assess vocabulary knowledge. We now focus in on the specific area of our dissertation. In this dissertation we investigate learning words by reading connected text – including extra vocabulary assistance – during computer-assisted oral reading. We focus on encountering a word for the first time, and on learning the meaning of a word.

1.5 Learning vocabulary from assisted oral reading

In this section we describe the process of learning vocabulary during assisted oral reading. We describe an informal model: a conceptual framework useful for identifying opportunities to improve vocabulary learning.

We can characterize how many words a student learns in a day of assisted oral reading as shown in Equation 1.1.

$$\boxed{\frac{\text{New words learned}}{\text{Day}} = \frac{\text{Time reading}}{\text{Day}} \times \frac{\text{Stories read}}{\text{Time reading}} \times \frac{\text{New words seen}}{\text{Story read}} \times \frac{\text{New words learned}}{\text{New words seen}}}$$

Equation 1.1. New words learned per day of assisted oral reading.

We define our thesis statement as follows, in the context of equation 1.1. We can help children learn vocabulary during assisted oral reading by (a) helping them encounter new words, and (b) helping them learn new words they encounter. We aim to help children encounter new words by increasing how much new material students read – not a guaranteed outcome when students have substantial control over their interaction with the software. We aim to help children learn new words they encounter by augmenting text to facilitate better learning than possible with the unaugmented text – not a guaranteed outcome since reading is already a reasonable way to build vocabulary. We verify each of these claims by empirical tests of modifications to the 1997-1998 version of Project LISTEN's Reading Tutor, a computer program that listens to children read aloud and helps them learn to read (Chapter 2).

The remainder of this dissertation is as follows. In Chapter 2, we present the 1997-98 baseline version of Project LISTEN's Reading Tutor. In Chapter 3, we describe how we modified the Reading Tutor to help children encounter new words. In Chapter 4, we describe how we modified the Reading Tutor to help children learn the meaning of new words. In Chapter 5, we compare the modified Reading Tutor against classroom instruction and one-on-one tutoring by certified teachers. (Classroom instruction is itself difficult to beat and represents the status quo; one-on-one tutoring is in many ways the "gold standard" for instruction – although expensive to provide.) Chapter 6 presents follow-on experiments to illuminate further directions for vocabulary help. Finally, we summarize our thesis contributions in Chapter 7. Appendices provide the materials we used for our experiments.

We now turn to describing the Reading Tutor. Afterwards, we will return to Equation 1.1 to discuss how each of the terms in Equation 1.1 plays out in detail in the Reading Tutor.

2 Project LISTEN’s Reading Tutor

This dissertation builds on a larger research project with years of history and publications: Project LISTEN. Project LISTEN’s Reading Tutor listens to children read aloud, and helps them learn to read. A detailed overview of the history of Project LISTEN lies outside the scope of this dissertation; see Mostow & Aist (FF 2001) for further information. Here we simply inform the reader of enough previous results to set our work in context.

1994 Reading Coach. A predecessor to the Reading Tutor, Project LISTEN’s Reading Coach provided assistance in oral reading (Mostow et al. 1994; see Mostow et al. 1993 for earlier work). In a 1994 study, 34 second graders comprehended a challenging third-grade passage 40% better with Reading Coach assistance than without (Mostow & Aist FF 2001), as measured by a comprehension test administered immediately after students had read the passages being tested. In that study, there was no assistive effect for an easier passage.

1996-1997 pilot study. Iterative redesign of the Reading Coach with concurrent usability testing resulted in the 1996 version of the Reading Tutor (Mostow, Hauptmann, and Roth UIST 1995, Mostow 1996 video). In a 1996 pilot study reported in Mostow and Aist (PUI 1997), 8 bottom 3rd graders at a low-income urban elementary school (Fort Pitt Elementary) used the 1996 Reading Tutor in a small room under individual supervision by a school aide. The six students who completed the study (one moved away; another was unavailable for post-testing) averaged a 2-year gain in eight months from pre-test to post-test on a school-administered Informal Reading Inventory. Mostow & Aist (FF 2001) provides details.

Summer 1997 Reading Clinic. During the summer of 1997, 62 students in grades K-5 used the Reading Tutor during a reading clinic at a low-income urban elementary school (Fort Pitt

Elementary). Concurrently, “the Reading Tutor underwent major design revisions of the “frame activities” – logging in and picking a story to read – to enable classroom-based use” (Mostow and Aist FF 2001).

1997-1998 formative and controlled studies. As Mostow and Aist report (FF 2001):

“During 1997-1998, students in 11 classrooms at an urban elementary school [Fort Pitt Elementary] used the Reading Tutor as part of a formative study to explore use of the Reading Tutor in a regular classroom setting. In Spring 1998, 63 students [completed the study – out of 72 who started – and] either read with the Reading Tutor, used commercial software, or received conventional instruction including other computer use. The Reading Tutor group gained significantly more than statistically matched classmates in the conventional instruction group on the Passage Comprehension subtest of the Woodcock Reading Mastery Test, even with only a fraction of the planned daily 20-minute sessions. No other significant differences were found” (Mostow and Aist FF 2001). The 1997-1998 study assessed Word Attack, Word Identification, Passage Comprehension, and fluency – but not Word Comprehension.

We used the 1997-1998 Reading Tutor used in the 1997-1998 studies as the baseline system for our dissertation. We based our research on oral reading for two reasons. First, reading is essential to building vocabulary. Second, adding instruction on the meaning of new words to oral reading allows spelling, receptive pronunciation, productive pronunciation, and meaning to be learned simultaneously. We carried out our research with Project LISTEN’s Reading Tutor for several reasons. First, we wanted to add vocabulary to the set of skills the Reading Tutor assisted with. Second, we wanted to exploit the use of the Reading Tutor as a research platform. Finally, the availability of students reading with the Reading Tutor meant that we could carry out experiments more efficiently than by developing separate software and educational implementations of such software for each experiment.

We now move on to describing the baseline 1997-1998 Reading Tutor.

2.1 Description of the baseline 1997-98 Reading Tutor

In this section, we describe the 1997-1998 Reading Tutor. We treated the 1997-1998 version as the baseline system for the experiments described in later chapters. We examine the baseline version from the outside in: social context, typical use, and technical details.

2.1.1 The Reading Tutor in its social context

The design of the 1997-1998 Reading Tutor focused on independent classroom use. There was a single Reading Tutor computer per classroom. Figure shows a student reading with the Reading Tutor while the teacher worked with the rest of the class.

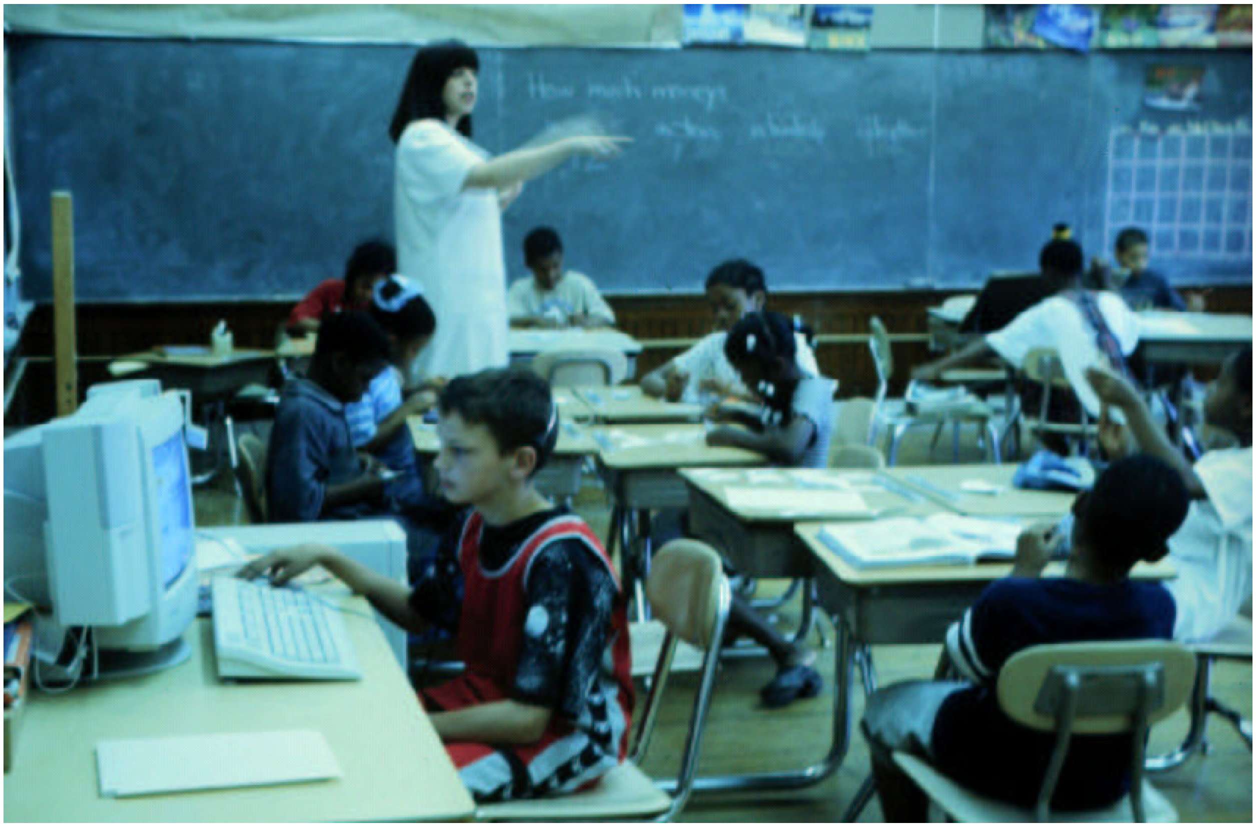


Figure 2.1. A student reads with the Reading Tutor while the teacher teaches the rest of the class.

Teachers are in the classroom to teach, not to tweak software. Although administrative functions in the 1997-1998 Reading Tutor permitted teachers to adjust the software if desired (through a special "teacher menu"), the Reading Tutor was designed to function with little or no direct usage by the teacher. Instead, students used the Reading Tutor independently, with guidance from the teacher. We term teachers' interaction with the software under this approach "indirect usage." Project LISTEN staff trained teachers and provided technical support, including on-site visits about once per week to collect recorded data and check each computer for problems.

Tutorial stories introduced students to various aspects of using the Reading Tutor.

Administrative functions allowed teachers to enroll new students, to add or edit stories, and to exit the software. In the formative study, teachers also set classroom policy for Reading Tutor usage, such as how many minutes each student should read per day, or how many stories each student should read per day. In the controlled study, we asked the teachers to implement the 20-25 minutes per day design.

The Reading Tutor recorded events in a user history database for later online use in making tutorial decisions. The Reading Tutor also wrote log files, recorded student speech, and saved speech recognizer output for research purposes. Project LISTEN staff first transferred the data to recordable CDs, Jaz™ disks, or other media. Then, they brought the data back to Carnegie Mellon. Finally, they archived the data for later analysis.

2.1.2 A prototypical session with the 1997-1998 Reading Tutor

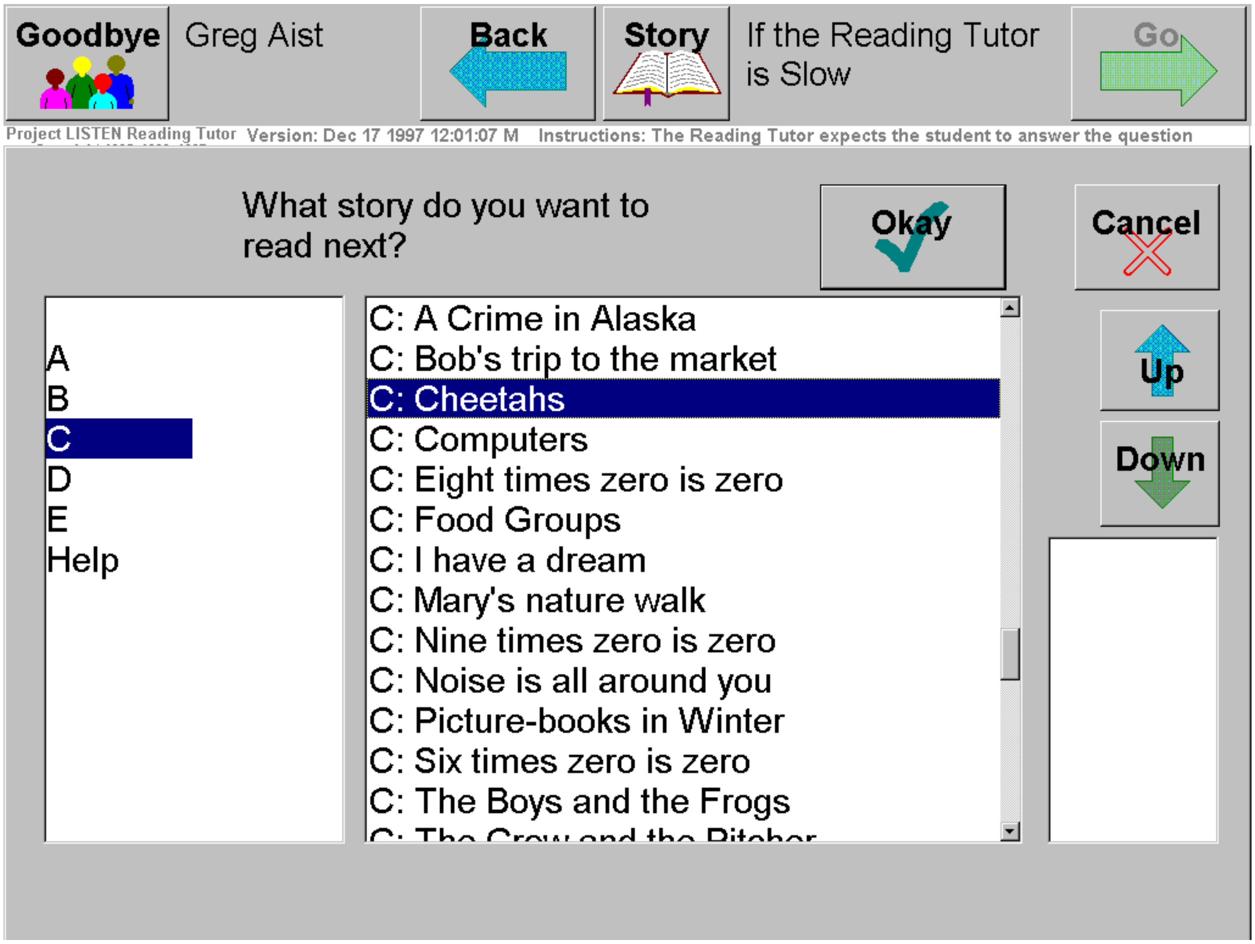
A prototypical student session with the 1997-98 Reading Tutor consisted of the following steps: log on, choose a story to read, read part or all of the story, (perhaps) choose and read more stories, and finally log off. We describe each step briefly.

2.1.2.1 Log on

Since the software took a few minutes to launch, we generally left it running during the day. Thus, the student did not need to launch the Reading Tutor. However, we needed to keep different children's data separate, to enable later analysis. Thus the student would begin a session by logging on. A student would **log on** by choosing his or her name from a list of enrolled students, and then confirming the chosen name by clicking on his or her birth month. (We used birth month as a lightweight password that the student would be unlikely to forget.)

2.1.2.2 Choose a story to read

Second, a student would **choose a story to read** from a list of all stories in the Reading Tutor. The 1997-1998 Reading Tutor displayed a list of all the stories in the system, and allowed the student to choose a story to read (Figure 2.2.). The Reading Tutor displayed the list of stories sorted first by level (K, A, B, C, D, E, Help) and then alphabetically by title. The Reading Tutor displayed the stories starting at the story level of the last story read, e.g. Level A in Figure 2.2. The Reading Tutor initially displayed the story list starting at the first story of the selected level, e.g. “A Crime in Alaska” in Figure 2.2. When a child clicked on a story level, the Reading Tutor scrolled to the first story of that level. When a child clicked on a story title, the Reading Tutor highlighted the title, read the title aloud, and – in order to guide the student to good voluntary choices – spoke an estimate of the story’s difficulty out loud. The estimate of difficulty was based on the percentage of words the student had not seen before, for example, “just about right” or “probably way too easy.” The child could then click *Okay* to begin reading the selected story.



Project LISTEN Reading Tutor Version: Dec 17 1997 12:01:07 M Instructions: The Reading Tutor expects the student to answer the question

What story do you want to read next?

A
B
C
D
E
Help

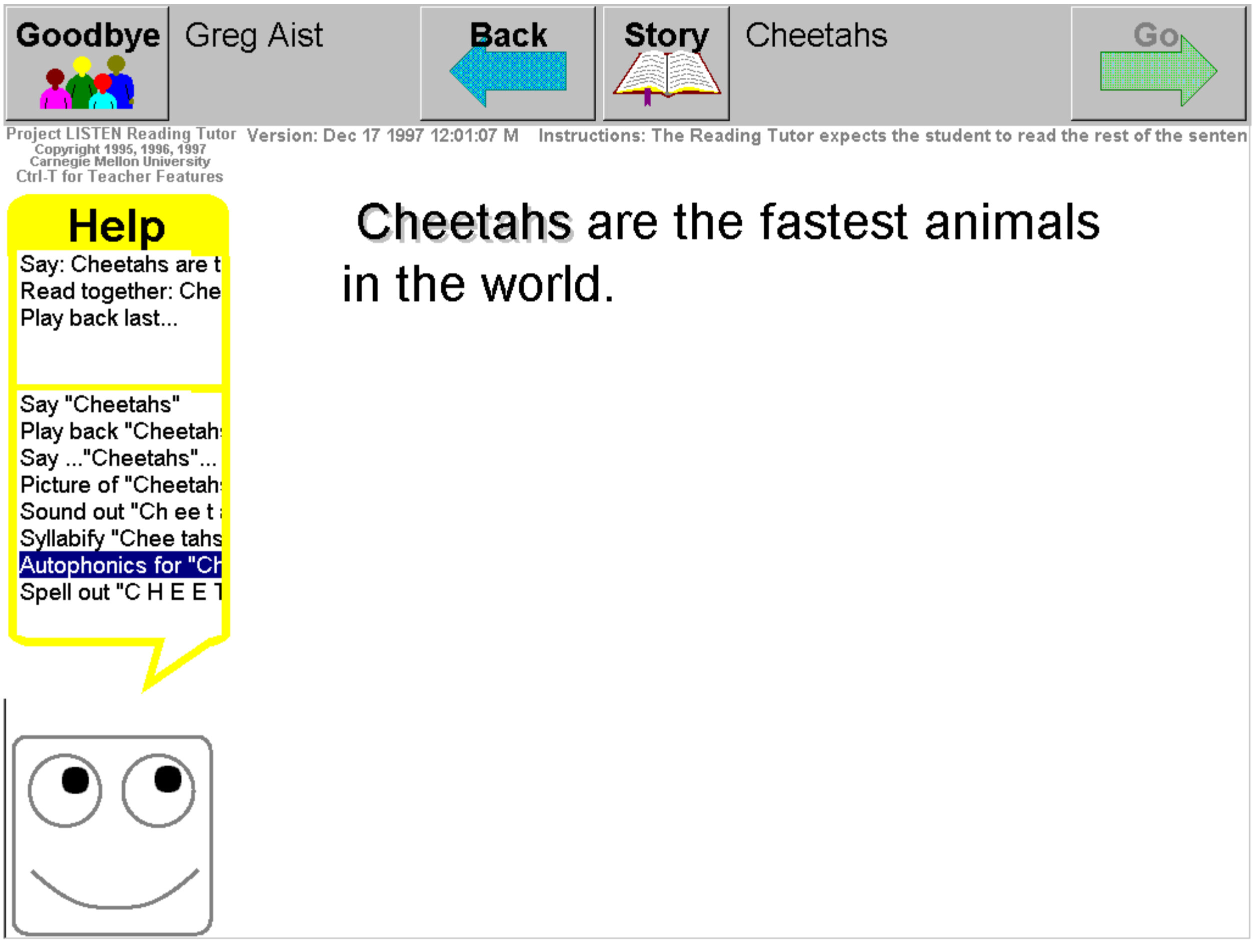
C: A Crime in Alaska
C: Bob's trip to the market
C: Cheetahs
C: Computers
C: Eight times zero is zero
C: Food Groups
C: I have a dream
C: Mary's nature walk
C: Nine times zero is zero
C: Noise is all around you
C: Picture-books in Winter
C: Six times zero is zero
C: The Boys and the Frogs
C: The Crow and the Pitcher

Figure 2.2. Story choice screen, 1997-1998 Reading Tutor. Level K stories were displayed with a blank level, " "; level A through E stories were displayed with a single-letter level; Help stories were displayed with the "Help" level.

2.1.2.3 Read all or part of the story aloud

After selecting a story, a student would **read all or part of the story aloud**. Basically, the Reading Tutor displayed one sentence at a time, listened to the student read aloud, and provided help on words it heard read incorrectly – words that the student may have missed or struggled with. The student could read a word aloud, read a sentence aloud, or read part of a sentence aloud.

Navigation. The student could click on *Back* to move to the previous sentence, on the face or on the items in the *Help* balloon to request help on the sentence, or *Go* to move to the next sentence (Figure 2.3). The Reading Tutor moved on to the next sentence when it had heard the student read every content word (Aist 1997 provides details). The student could click on *Story* to pick a different story (1997-1998 version only), or on *Goodbye* to log out.



Version: Dec 17 1997 12:01:07 M

Instructions: The Reading Tutor expects the student to read the rest of the senten

Cheetahs are the fastest animals in the world.

Figure 2.3. Reading a story in the 1997-1998 Reading Tutor.

Reading assistance. The Reading Tutor responded when it heard mistakes or when the student clicked for help. Responses were constructed by playing hints or other help in recorded human voices. The help that the Reading Tutor provided sought to balance the student’s immediate goal of reading the word or sentence with the longer-term goal of helping the student learn to read

(Aist and Mostow CALL 1997, Mostow and Aist CALICO 1999). Help included:

1. Read the entire sentence using a recording of a human narrator's fluent reading, to model correct reading. While playing the (continuous) recording, the Reading Tutor would highlight each word as it was spoken, which we call word-by-word highlighting.
2. Read the entire sentence by playing back isolated recordings of a single word at a time, in order to allow students to hear one word read at a time. Because these recordings may be in different voices, we call word-by-word playback "ransom note" help.
3. Recue a word by playing an excerpt from the sentence narration of the words leading up to that word (along with word-by-word highlighting), in order to prompt the student to try (re-) reading the word. For example: If the text is **Jack and Jill went up the hill to fetch a pail of water**, the Reading Tutor could recue **hill** by first reading **Jack and Jill went up the** out loud, and then underlining the word **hill** to prompt the student to read it.
4. Give a rhyming hint that matches both the sound (phoneme sequence) and the letters (grapheme sequence) of the target word, in order to give a hint on how to read the target word, and to expose the student to related words. For example, if the word is **hill**, give the word **fill** as a spoken and displayed rhyming hint, but not the word **nil** because its spelling does not match.
5. Decompose a word, syllable-by-syllable or phoneme-by-phoneme, to model the process of sounding out words and to call attention to letter-to-sound mappings. For example, say /h/ while highlighting **h**, then say /i/ while highlighting **i**, then say /l/ while highlighting **ll**.
6. Show a picture for a word, in order to demonstrate word meaning and to increase engagement. For example, if the word is **apple**, show a drawing of an apple. Fewer than 200

words had pictures in the 1997-1998 version.

7. Play a sound effect, perhaps to demonstrate word meaning but primarily to increase engagement. For example, if the word is **lion**, play the roar of a lion. Fewer than 50 words had sound effects in the 1997-98 version; most were names of animals, such as *seagulls*, *tiger*, and *wolf*.

2.1.2.4 (Perhaps) choose and read more stories

The student could pick a new story at any time by clicking the *Story* button. When the student finished reading a story, the Reading Tutor would then display the story choice menu (Figure 2.3) again.

2.1.2.5 Log out

When the student clicked *Goodbye*, or when a long period of time had elapsed without any student activity, the Reading Tutor would log the student out and wait for the next student. Logging the student out based on timeout was intended to keep students' data separate from their peers' data; students did not always log themselves out.

Now that we have described a typical session with the 1997-1998 Reading Tutor, we briefly discuss its technical aspects.

2.1.3 The Reading Tutor as a software program

We describe the Reading Tutor as a software program: first its inputs and outputs, and then its hardware requirements.

The 1997-1998 Reading Tutor received as input the following: the student's speech from a noise-canceling headset or handset microphone; mouse clicks on on-screen buttons; and keyboard entry when writing stories or enrolling new students. Mostow et al. (1994) described how the Reading Tutor adapted the Sphinx-II speech recognizer (Huang et al. 1993) to listen to children read aloud. The Reading Tutor tracked the student's position in the sentence, flagged (presumed) mistakes, and decided when the student was stuck or needed help to continue. The 1997-98 Reading Tutor displayed text, graphics, and pictures on a standard computer monitor, and played sound through computer speakers or headphones.

The 1997-98 Reading Tutor was a software program written primarily in C++, with small sections written in perl, SQL, and MS-DOS batch language. The 1997-98 Reading Tutor ran under Windows NT 4.0 on a standard commercially available IBM-compatible computer with 128 megabytes (MB) of memory, costing approximately $2,000 at the time. So that it could talk and listen at the same time, the 1997-1998 Reading Tutor required a full-duplex sound card and device driver. The 1997-1998 Reading Tutor also required a 200 MHz or faster Pentium™ processor and 1024x768 screen resolution. Installation was from compact disc (CD-ROM), and data was stored to hard disk and collected later using recordable CDs on a compact disc writer.

We now compare the Reading Tutor to other reading software that uses speech recognition

2.2 Comparison of baseline 1997-1998 Reading Tutor and other software

To place the baseline Reading Tutor in its research context, and clarify its differences with respect to similar software, we compare it here to other software. We focus on software that (a) helps with reading, (b) in a child's first language, (c) using speech recognition. Readers who are interested in software outside these constraints may refer to (Aist SR-CALL 1999) for an overview of speech recognition in second language learning, and (Schacter 1999) for an overview of conventional and software-based reading instruction for a child’s first language. Whines (1999) provides a detailed comparison of some of the systems described below.

The **Speech Training Aid (STAR)** developed by DRA Malvern adapted automatic speech recognition to help children practice reading single isolated words (Russell et al. 1996). The 1997-98 Reading Tutor listened to children read connected, authentic text.

Talking and Listening Books, also described by Russell et al. (1996), used continuous text but employed word spotting techniques to listen for a single word at a time.

Let’s Go Read (Edmark 1997) incorporated speech recognition into a variety of single-phoneme and single-word exercises. The 1997-1998 Reading Tutor focused on assisted reading of authentic text.

Watch Me! Read (IBM 1998, Williams et al. 2000) adapted speech recognition to teach reading from continuous text, but took a traditional talking-book approach using trade books with lots of pictures but small amounts of text in small fonts. The 1997-1998 Reading Tutor used child-friendly large fonts and placed primary emphasis on reading text – not on looking at pretty pictures.

Now that we have described the 1997-1998 baseline Reading Tutor and compared it to other reading software that uses speech recognition, we revisit how children learn words in computer-assisted oral reading. We reconsider Equation 1.1 in the context of the 1997-1998 Reading Tutor and subsequent improvements.

2.3 Learning vocabulary in the Reading Tutor

How many words can we expect students to learn from the Reading Tutor? We can conceptualize this problem as a specialization of Equation 1.1, as follows (Equation 2.1).

$$\frac{\text{New words learned on RT}}{\text{Day}} = \frac{\text{Time on RT}}{\text{Day}} \times \frac{\text{Stories read on RT}}{\text{Time on RT}} \times \frac{\text{New words seen on RT}}{\text{Story read on RT}} \times \frac{\text{New words learned on RT}}{\text{New words seen on RT}}$$

Equation 2.1. Words learned per day on the Reading Tutor (RT).

We can split the reading that a student does into two categories: (a) reading with the Reading Tutor, and (b) everything else (outside the scope of this dissertation). Thus from here on out we will generally omit “on RT” from our equations, as we focus on just the learning of new words that takes place in the Reading Tutor.

In the case of reading with the Reading Tutor, “how much reading” translates into how many days a student has a session with the computer, and how many minutes each session lasts. How often the Reading Tutor gets used by whom for how long depends on who sets policy for Reading Tutor use, and in any event lies outside the scope of this thesis. Therefore, for the purposes of the present discussion we will take the number of days allocated for Reading Tutor use per year as externally determined, and likewise we consider the number of minutes of Reading Tutor use per day as also externally determined. How frequently we expect students to read with the Reading Tutor, and for how long each session, have in practice varied for different studies and in different contexts of use.

In order to identify areas for improvement in computer-assisted oral reading, we decompose every term in this equation in terms of how each aspect was implemented in the 1997-98 Reading Tutor and in the modifications we made for this dissertation. A new software design or educational implementation would add new subterms. For example, adding a feature to inject new text into an old story would change the composition of the term “new words seen per story read.”

2.3.1 Stories read per unit time

$$\frac{\text{Stories read}}{\text{Time per day on Tutor}} = \frac{\text{Time reading}}{\text{Time per day on Tutor}} \times \frac{\text{Words}}{\text{Time reading}} \div \frac{\text{Words}}{\text{Story read}}$$

Equation 2.2. Stories read per time on Tutor.

Assuming that the number of minutes per session is externally determined, we can represent the number of stories read per day as shown in Equation 2.2.

We now discuss each of the terms in Equation 2.2 in turn.

Time reading / Time per day on Tutor is whatever percentage of time in a session is dedicated to reading stories, instead of other activities such as logging in, picking stories, writing stories, or answering comprehension questions.

Words / Time reading is essentially reading rate. In the Reading Tutor, however, this rate will be decreased by the time needed for the Reading Tutor to provide help, as well as system delay in showing the next sentence or in responding promptly. In addition – as with independent reading – brief distractions may decrease the student's reading rate even if he or she nominally remains on task. Finally, rereading a story will be faster than reading it for the first time.

Words / Story read is story length. Harder stories will of course tend to have more words.

2.3.2 New words seen per story read

$$\frac{\text{New words}}{\text{Story}} = \text{New words in original text} + \text{New words in additional text}$$

Equation 2.3. New words seen per story read.

Equation 2.3 describes how many new words a student will see in a story, for both the original text and for any text added by the Reading Tutor (see below).

New words in original text is zero for a previously completed story, and varies with story difficulty for a previously unread story.

New words in additional text is zero if the Reading Tutor does not add any text to the original story, but may be greater than zero if (for example) the Reading Tutor introduces a comparison between a word in the text and some other word that the student has not seen before.

2.3.3 New words learned per word seen

$$\boxed{\frac{\text{New words learned}}{\text{New words seen}} = \frac{\text{New words actually read}}{\text{New words seen}} \times \frac{\text{New words learned}}{\text{New words actually read}}}$$

Equation 2.4. New words learned per new word seen.

Equation 2.4 describes the ratio of new words learned to new words seen – equivalently, the probability that a student learns a new word if he or she sees it.

New words actually read / New words seen. A student must actually read the word – not just skip over it – to learn its meaning. A poor reader might be tempted to skip over a difficult word rather than tackle the word head-on.

New words learned / New words actually read. A student of high reading ability will learn more from context than a student of low reading ability (McKeown 1985, Shefelbine 1990). Not all contexts are equal for word learning (Beck, McKeown, and McCaslin 1983). A sentence with rich semantic content will help more than a list of words. Adding useful information to text might increase the ratio of new words learned to new words actually read.

2.4 Goal: Help students encounter new words, and learn them

We focus in this dissertation on the last two factors in equation 2.1: new words seen per story, and new words learned per word seen. First, students must encounter new words. Second, they must learn the meaning of new words when they initially encounter them. We modified Project LISTEN's Reading Tutor to be more effective at each of these tasks.

We next present the improvements we made to the Reading Tutor, along with experiments evaluating their effectiveness. Chapter 3 presents improvements to the Reading Tutor's story choice policy and Chapter 4 presents experiments on providing vocabulary help.

3 Improving story choice

The tale of Reading Tutor story choice is one of finding a balance between students' interests and the Reading Tutor's educational goals. Children have their own agenda when using software, which may or may not match the desired educational outcome. As Hanna et al. say, "When analyzing usage by children, we look at the goals of the product and the goals of children. The goal of the product may be to teach the alphabet, but children will probably not play with the product because they want to learn the alphabet. A child's goal may be to explore and find out what happens or to win a game" (Hanna, Risden, Czerwinsky, and Alexander 1999). The Reading Tutor's goal is to help students learn to read. A student's goal may be reading a particular story, writing a story, exploring the story menus, or something else.

In the 1997-1998 school year, students used the Reading Tutor in a classroom independently. To allow students to read stories that interested them, and to increase students' interest in what they are reading by maximizing learner control, the 1997-98 Reading Tutor allowed students free choice of any story on the Reading Tutor. Stories available included non-fiction, poems, and fictional narratives. We observed a number of problems with story choice. First, teachers reported that some students chose to read the same story over and over again – whether a favorite story such as an excerpt from Martin Luther King's "I have a dream" speech, or an easy story such as the nursery rhyme "Jack and Jill went up the hill." In order to encounter new words in natural in-context use, students must read text they have not read before. Second, we observed that some students consistently chose stories that were too easy. To learn new words, students must read text hard enough to contain new words. We wanted to improve story choices to get children to see more new words than they would on their own – where new words is new words seen per story read, in Equation 1.1 and Equation 2.1.

In this chapter, we discuss improvements we made to the Reading Tutor’s story choice policy. First, we discuss design considerations; next, we discuss how we implemented our revised policy; then, we evaluate the results; finally, we discuss lessons learned.

3.1 Revising the story choice policy

How can we get students to read new and challenging material? In this section, we first enumerate design considerations for a revised story choice policy. Next, we describe some of the options we considered. Then, we reveal the revised policy we chose to implement.

3.1.1 Design considerations for revised story choice policy

During the process of constructing and analyzing possible story choice policies, we made explicit some desired characteristics for story choice. A good story choice policy is:

Classroom-compatible. Does the policy support teachers’ indirect use of the Reading Tutor, by allowing teachers to direct students towards particular stories? Does the policy encourage healthy competition – such as for new stories read – but not encourage useless competition such as the number of times the student read their favorite story? Does the policy enable students to read stories that their friends recommend? A good story choice policy will reflect the realities and opportunities of classroom use.

Usable. Does the policy present clear and unambiguous choices to the student? Is the policy simple to describe to students and teachers? Confusing or complicated behavior can cause loss of face validity.

Acceptable. Does the policy result in choosing stories that students actually read? If students boycott stories chosen by the computer, the Reading Tutor will be less effective.

Efficient. Does the policy take at most as much time as the student-only story choice policy that the 1997-1998 Reading Tutor used? Every minute spent choosing a story is a minute not spent reading a story. A revised story choice policy should at least take no more time than the student-only story choice policy of 1997-1998.

Effective. Does the policy guarantee that every student will read some new material regularly? Reading new material is necessary for encountering new words in context. Does the policy permit (some) re-reading of material? Successful reading programs also include some re-reading (Schacter 1999).

3.1.2 Story choice policy: Options considered

In order to improve the shared reading process, we wanted the Reading Tutor to do better at helping children pick appropriate material. One might imagine that each student had his or her own list of stories to read, and that the teacher or some administrator would adjust an individual student's reading list. However, we have found that teachers want to teach, not tweak software; they seem to prefer interacting with the Reading Tutor indirectly by guiding students towards productive behaviors and (occasionally) checking on student activity or progress. For example, one teacher put a list of stories to read on a 3x5 index card and set the card on top of the Reading Tutor monitor. We wanted to support such indirect teacher use under the revised story choice regime – but not require it. Thus we needed to devise a story choice policy that would be robust with no direct teacher involvement whatsoever, but permit indirect teacher involvement as did the original policy.

Choosing what to read allows children control and lets them select stories that interest them. Therefore, we did not want to take away student choice altogether. Some of the less drastic options we considered were:

Sorting the story list. We could arrange the list of stories so that students were more likely to choose new stories. For example, we could put new stories first on the list. However, any policy that merely sorted the list of stories would not be guaranteed to be effective. Why? A student could simply ignore the suggestion and choose the same story over and over again.

Restrict the story list. We could restrict the list of stories that students could choose from. However, any non-trivial restriction could be difficult for children to understand. Also, how would the student get to choose an old favorite some of the time without being able to choose it all of the time? Finally, students might lose the opportunity to choose stories that their friends recommended.

Provide different lists. We could have the student alternately choose from two separate lists of stories. For example, the Reading Tutor could sometimes show a list of new stories, and other times a list of old stories. Using different lists is simple to describe, should take no more time than choosing from a single list, seems fair because the student gets to make all of the (restricted) choices, guarantees that students read some new material, and lets the student re-read favorite stories. Using different lists would be easy to implement, but might be confusing. The student would be required to make choices from similar-looking lists of stories, but the actual material available on each list would be completely different (old, and new). Good design might clarify which list showed old stories, and which list showed new stories – e.g. labeling lists “Old” and “New”, and using different graphical appearances for the two lists. Nonetheless, a more substantive objection remains: why require the student to pick an old story if he or she would rather read something new? Re-reading is beneficial, but why limit the amount of challenge students could seek for themselves?

Take turns. We could have the student choose some of the stories, and the Reading Tutor choose some of the stories, with the Reading Tutor always choosing new stories. Take Turns would let the teacher influence the student’s choices, and would let students choose to read stories their friends recommend (at least when it is the student’s turn to choose.) Taking turns would be easy to describe to students and teachers, and (sometimes) let the student choose an old favorite to re-read. Would it be easy to use? Each decision the student would make is from the same stable list of stories, which should be simpler to navigate than two different, changing lists of stories. Would taking turns be fair? The student doesn’t get all the choices – but the choices he or she does get are not restricted. Would taking turns be quick? If student choice were to take the same amount of time as with the student-only story choice policy, and Reading Tutor choice were to take negligible time, this policy should take about half the time of the 1997-1998 story choice policy. Finally, taking turns choosing stories would guarantee that every student would read some new stories.

Of these general policies, providing different lists and taking turns offered the most advantages and posed the fewest drawbacks. We decided to implement a policy of taking turns. We believed that taking turns would be slightly easier to explain than presenting different lists. We also liked that taking turns (sometimes) would provide the student with an unfiltered free choice of what to read. Finally, taking turns should take less time than having the student choose from alternating lists, giving the student more time to read stories.

Take Turns yields an interesting side benefit: since the Reading Tutor gets to pick stories half of the time, it can use its choice for many different purposes. (Introductory stories, as described in Section 3.2.1, are one example.) For 1999-2000, we programmed the Reading Tutor to choose new stories; in later work (not described in this dissertation) we used the Reading Tutor’s turn to conduct an experiment. We could also use the Reading Tutor’s turn to introduce other activities such as a passage to assess fluency.

Accordingly, we introduced a Take Turns story choice policy for the 1999-2000 Reading

Tutor.

3.2 Implementing Take Turns

Take Turns consisted of three components:

1. A mechanism to let the Reading Tutor and the student take turns choosing stories, instead of always allowing the student to choose.
2. A mechanism for the Reading Tutor to pick stories.
3. A mechanism for allowing students to choose stories, with modifications aimed at encouraging students to choose appropriately challenging material.

We now discuss each of these components in more detail.

3.2.1 Reading Tutor and student take turns choosing stories

What did Take Turns need to do? Obviously our implementation of Take Turns had to let the Reading Tutor and the student take turns picking stories. The Take Turns algorithm had to allow for an initial tutorial introducing the Reading Tutor, as described below. Furthermore, the Take Turns algorithm had to be robust, despite software crashes, students' attempts at circumventing the story choice policy to escape the Reading Tutor's turn, and varying classroom practices as to how many stories to read in a day.

We can sum up the algorithm as follows: "Every day, decide randomly whether the student or the Reading Tutor chooses the first story to read. Then, take turns for the rest of the day." Table 3.1 describes both components of the story choice algorithm: what happened when it was time to pick a story, and what happened when the student finished reading a story.

When it is time to pick a story…	
If tutorial not finished…	Then, choose the tutorial – a prespecified story introducing the Reading Tutor.
Otherwise, if this is the first time this student has logged in today…	Then: Pick randomly (50/50) who chooses next story, Reading Tutor or student. Set *who_chooses* to result.
Otherwise …	Use value of *who_chooses*: if *who_chooses* = student, student chooses story; else *who_chooses* = Reading Tutor, so Reading Tutor chooses story.
When the student has just finished reading a story…	
If *who_chooses* = Reading Tutor	Then, set *who_chooses* = student.
Otherwise, if *who_chooses* = student	Then, set *who_chooses* = Reading Tutor.

Table 3.1. How to Take Turns choosing stories.

We comment on a few aspects of Take Turns.

Tutorial. We wanted to reduce variation in student behavior due to differences in initial training. To give students a uniform introduction to the Reading Tutor, we included a prespecified tutorial story that described how to operate the Reading Tutor. When a student logged in, the Reading Tutor first checked to see if the student had completed the tutorial story. If the student had not completed the tutorial, the Reading Tutor selected the tutorial story. However, the 1999-2000 version of the tutorial proved too difficult for young readers, and was therefore replaced for the 1999-2000 study by initial small-group training given by a certified teacher on Project LISTEN's staff. More recent interactive tutorials (e.g. Summer 2000) were developed for younger children; a full description is outside the scope of this dissertation.

Who picks first on a particular day. The decision of who got to choose the first story of the day was randomized instead of systematic so that if a particular teacher's policy were to have each student read only one story, students would still read both Reading Tutor-chosen stories and stories they chose themselves. The result of the random decision was stored on disk, so as to survive crashes and software restarts.

Resetting *who_chooses* at the end of a story. By making *who_chooses* persist until either the student finishes a story or at least one day has passed, we aimed to prevent students from "cheating" – getting around the Reading Tutor's story choice by logging out and then logging back in again.

3.2.2 Reading Tutor story choice

The 1999-2000 Reading Tutor assigned each student to a recommended reading level based on the student's age, and adjusted the recommended reading level based on the student's performance. The Reading Tutor tried to pick new stories at the student's recommended reading level. If no story was available at the recommended reading level, the Reading Tutor chose a harder story.

Initial assignment. The Reading Tutor assigned the student to an initial recommended reading level based on age. The student's age was derived from his or her birthday month and day, and initial age at time of enrollment. Initial assignments were as follows (Table 3.2).

Student's age	Initial recommended reading level	Story level grade equivalent
7 years or younger	K	Kindergarten
8-9 years old	A	First grade
10-11 years old	B	Second grade
12-13 years old	C	Third grade
13 years old or older	D	Fourth grade

Table 3.2. Initial assignment to recommended reading level.

The initial level was deliberately low to avoid frustrating poor readers, and relied on level adjustment to quickly reach story levels that would challenge the better students.

Level adjustment. The Reading Tutor adjusted the student's recommended reading level in order to keep difficulty reasonable. Reading researchers commonly consider three levels of text difficulty: independent reading level, instructional level, and frustration level. A common criterion for determining reading level is word reading accuracy; for example, “A book is said to be at a child’s independent level if 95 – 100 percent of the words can be read correctly. Instructional level books can be read with a 90 – 94 percent level of accuracy. Frustration level reading involves text read by a child at the 89 percent accuracy level or below” (California Department of Education 1996).

How can we measure difficulty? Reading fluency is highly correlated with comprehension (Pinnell et al. 1995). Using speech recognition to listen to children read aloud enables direct, unobtrusive measurement of oral reading performance (for example, see Mostow and Aist AAAI 1997). However, identifying students’ errors on individual words by machine remains imperfect. Thus, we based the level adjustment criterion for the Fall 1999 version not on the student’s oral reading accuracy, but on the student’s assisted reading rate: the number of words accepted by the Reading Tutor per minute of assisted reading. This rate is distinct from, but reflects, the student’s oral reading fluency.

The level adjustment policy, described in Table 3.3, applied only after the student had finished a new story. The teacher could theoretically override the recommended reading level if necessary, by using an administrative mode of the Reading Tutor – but in practice did not. In retrospect, story-level-specific promotion and demotion thresholds would have been an improvement over a level-independent threshold, because reading rates are expected to increase with grade level.

Last story completed, compared to recommended reading level	Accepted words per minute (AWPM) < 10	10 ≤ AWPM < 30	AWPM ≥ 30
Easier	Move down one level	Stay same	Stay same
At level	Move down one level	Stay same	Move up one level
Harder	Stay same	Stay same	Move up one level
Noncomparable	Stay same	Stay same	Stay same

Table 3.3. Level adjustment policy.

Sorting the stories. The Reading Tutor sorted the stories by level as follows. Each story in the Reading Tutor had a story level, assigned by Project LISTEN team members. Levels K, A, B, C, D, and E were used for kindergarten through fifth grade respectively, rather than the (possibly stigmatizing) K, 1-5. The Help level was used for help stories. (Help stories were about how to operate the Reading Tutor and handle common technical problems.) Student-authored stories were assigned to level U for “unleveled.” The story levels K through E were ordered by increasing difficulty, but levels U and Help did not have a defined difficulty relationship with any other level. The Reading Tutor sorted all of the approximately 200 stories into five categories, as shown in Figure 3.1: 1. previously read stories, 2. new stories below the student’s recommended reading level, 3. new stories at the student’s recommended reading level, 4. new stories above the student’s recommended reading level, and 5. stories at a noncomparable level.

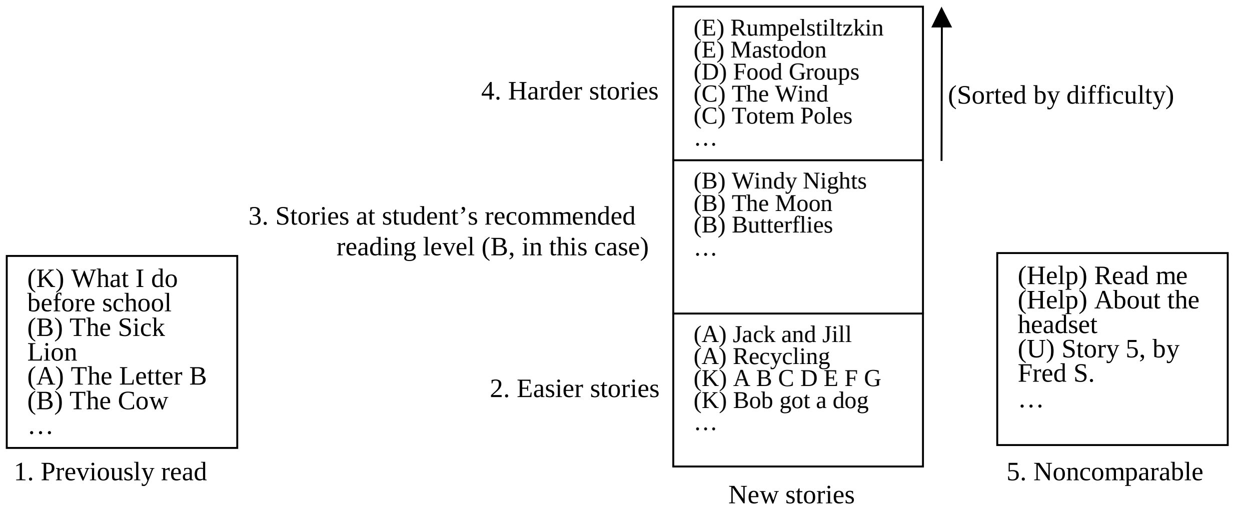


Figure 3.1. Sorting stories into levels for Reading Tutor story choice.

The Reading Tutor first tried to choose an unread story at the student's recommended reading level. If a student had completed all the stories at the recommended reading level, the Fall 1999 version chose randomly from unread stories at *all* of the harder levels, not just the next highest level. This unintended behavior meant that students were therefore sometimes faced with a story that was too difficult, a problem that showed up only after extended use had enabled some students to finish all the stories at one level. For Spring 2000, we modified Take Turns so that the Reading Tutor would sort the stories by level and choose randomly from the stories at the next level immediately above the student's recommended reading level.

3.2.3 Student story choice

When it was the student's turn to choose a story, the student was free to choose any Reading Tutor story to read. The student could alternatively choose to write and (optionally) narrate a story.

We wanted to help the student make good story choices. Specifically, we wanted to:

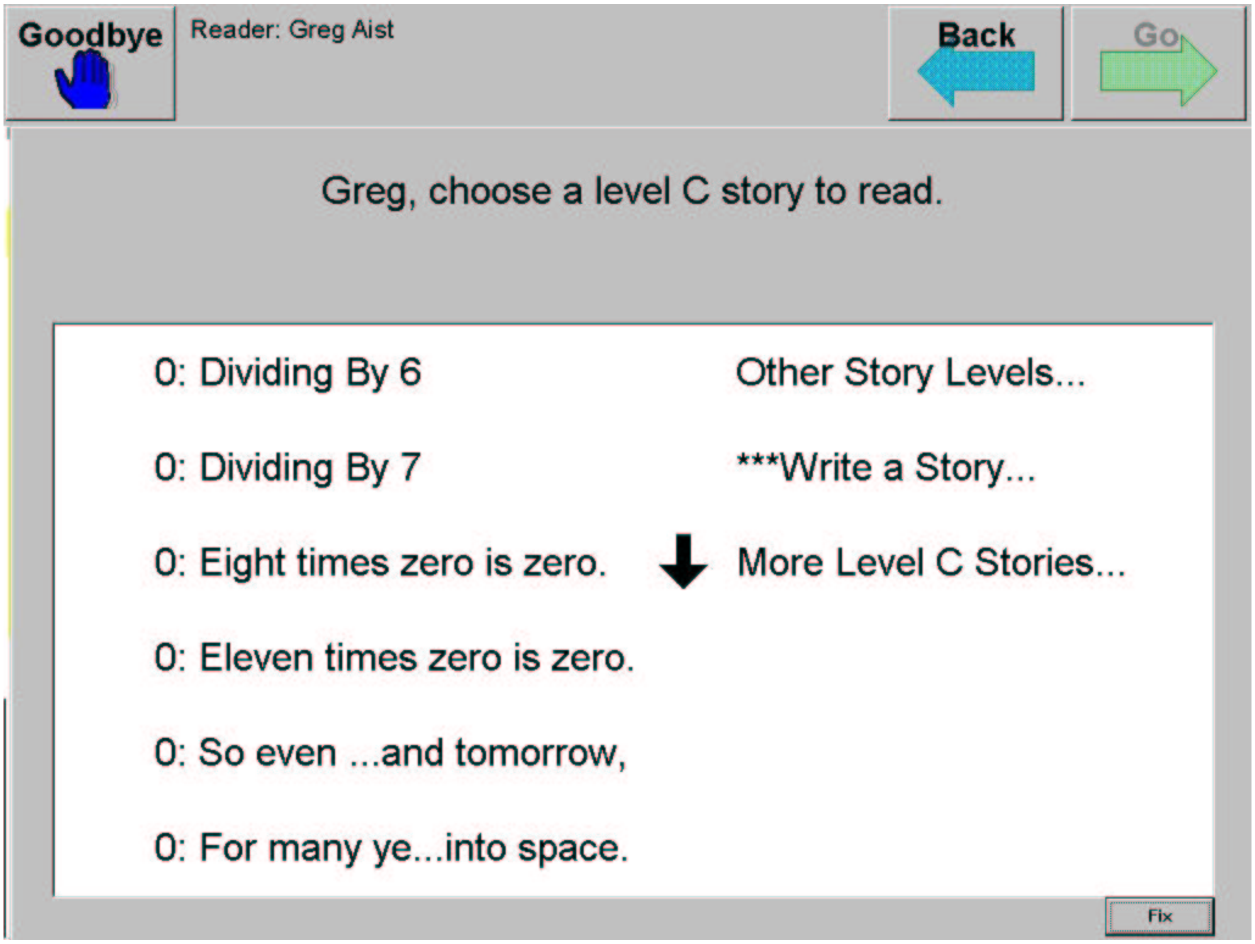
Encourage students to read new stories. To guide the student to new stories, and to provide a visible record of what the student read, we modified the story choice menu to show the number of times the student had read each story. We were concerned that students would use "how many times I read the story Life in Space" as a score – so after the first few readings the Reading Tutor just displayed ">4" for the number of readings.

Provide a varied menu of stories to choose from. The number of stories available in the Reading Tutor was far larger than the number of story titles that would fit on the screen in a child-friendly (large) font. We were concerned that some students might follow the path of least effort and always pick from the first few story titles shown on the screen. Therefore, we modified the story choice menu to start at a random screenful in the (alphabetically) ordered list of stories, rather than always at the top.

Support finding a favorite story. One important aspect of reading is reading stories you like. We wanted students to be able to find a favorite story if they wanted. Thus we kept the list in order rather than scrambling it, because scrambling would have made it harder to find a particular story.

For 1999-2000, we simplified the menu interaction. The 1997-1998 Reading Tutor required at least two clicks for story choice: one click to select a story and another click to confirm the selection. The 1999-2000 Reading Tutor required only one click to select a story, but let the student click on the *Back* button to return to the story choice screen and choose again. The change to "one-click pick" is described in more detail in (Aist & Mostow ITS-PA 2000). This redesign aimed to accommodate a wide range of students. For example, reading the story titles out loud to the student supported non-readers who couldn't read the titles. More advanced readers, who could read the titles without help, could simply click on a title at any time.

Figure 3.2 shows the revised story choice screen. To choose a story, the student could click on a title at any time. The Reading Tutor spoke the prompt displayed at the top (here, “Greg, choose a level C story to read”) and then read each title out loud while highlighting it in yellow. If the student did not click on a story title, the Reading Tutor would read the list of titles again, and eventually time out – logging the student out. Clicking on the item “More Level __ Stories” showed more stories at the current level, “More Level C Stories” in this example.



Reader: Greg Aist

Greg, choose a level C story to read.

0: Dividing By 6

0: Dividing By 7

0: Eight times zero is zero.

0: Eleven times zero is zero.

0: So even ...and tomorrow,

0: For many ye...into space.

Other Story Levels...

***Write a Story...

More Level C Stories...

Figure 3.2. Story choice screen, fall 1999. The story “Life in Space” is shown here with a title abbreviated from its first sentence, “For many years, the United States and Russia worked separately on going into space.” The number of times the student has read each story is displayed in front of the story title.

3.3 Evaluation of Take Turns story choice policy

We evaluated Take Turns using the criteria specified previously in Section 3.1.1: classroom-compatibility, usability, acceptance, efficiency, and effectiveness.

3.3.1 Classroom-compatibility

By design, Take Turns accommodated teachers' recommending stories to children, and allowed children to pick stories their friends recommended. Take Turns also permitted children to re-read favorite stories. These functions were supplemented by displaying on the Hello screen, which was displayed when no student was logged in to the Reading Tutor, various information about the students' use of the software. For example, in order to encourage (healthy) competition, the "Hello" screen showed how many minutes each student had read that day, and how many new words and new stories each student had seen.

3.3.2 Usability

We tested the Take Turns story choice policy on several occasions for usability and acceptance. First, we tested a preliminary version at Fort Pitt Elementary School in several one-day informal trials during the spring of 1999. Second, several children at the CHIkids program at the 1999 Conference on Computer-Human Interaction (Pittsburgh) used the Reading Tutor with the Take Turns story choice policy over a period of several days. Third, over fifty students at Fort Pitt's 1999 Summer Reading Clinic used the Reading Tutor with the Take Turns story choice policy.

These tests exposed a number of concerns. First, some students were frustrated when the Reading Tutor sometimes chose material that was too hard for them. We tuned the reading level adjustment thresholds by reading ourselves as if to simulate good and poor reading, and by checking students’ assisted reading rates during Reading Tutor use. While not necessarily perfect, the new thresholds were good enough that students tolerated the Reading Tutor’s choices, as described later.

Second, as we expected, some students were frustrated because they wanted to read a different story than the one the Reading Tutor picked. As deployed, the Reading Tutor aimed to ameliorate student frustration by explaining its choices with phrases like “Read this story, and then YOU’LL choose one.” We suspect that such explanation may have helped, but did not totally remove frustration.

Third, the Reading Tutor would sometimes pick a story that was actually the middle of a story in multiple parts. For example, *Cinderella* was split into several parts to make each part short enough for a student to read in a single session. Project LISTEN programmer Andrew Cuneo made each part of a multipart story an explicit prerequisite for the next part, and modified the Reading Tutor to choose only stories whose prerequisite (if any) the student had finished reading. What had been common sense for children – read sequential stories in order – had to be explicitly programmed into the software.

3.3.3 Acceptance

We now turn to considering acceptance. We looked at several aspects of story choices: story completion, writing stories, story levels, and re-reading stories.

3.3.3.1 Story completion

To measure students’ acceptance of their own and Reading Tutor story choices, we first looked at how often students finished stories that they started reading. Using the data recorded by the

Reading Tutor, we calculated how many stories were started vs. how many were finished.[1] We examined how often stories started were actually finished.

Table 3.4 shows the per-student average number of stories started and finished for the Fall 1999 Centennial data, organized by new vs. old story and who chose the story (Reading Tutor or student). To calculate the per-student average, we averaged over all students, thus weighting students equally instead of skewing the average towards those who read more stories by simply aggregating over all stories.

	Reading Tutor	Student – new stories	Student – old stories
Total number of stories	1509 finished out of 5632 started	540 finished out of 1206 started	852 finished out of 1334 started
Percentage of stories finished: per-student mean and standard deviation	39.2% ± 20.8%	52.9% ± 23.2%	61.2%± 23.7%

Table 3.4. Stories finished out of stories chosen, Fall 1999 data.

The percentage of stories finished was lower in the Reading Tutor chosen stories, at 39%. The difference between Reading Tutor-chosen stories and student-chosen stories was perhaps due to the Reading Tutor's choice of harder stories, which tended to be longer.

Second, did students eventually finish *something* the Reading Tutor chose for them? Here, clearly yes; on average, students finished a larger number of new stories that the Reading Tutor chose than the number of new stories finished that they chose themselves (Figures 3.3 and 3.4). A substantial portion of students' choices were dedicated to re-reading old material – but that's the point: the Reading Tutor's choices were aimed at increasing the amount of new material read. So students were eventually accepting *some* Reading Tutor selections – enough to finish a story!

[1] On one occasion in Fall 1999, the Reading Tutor failed to record that a student had opened a story because the disk was full (Tuesday, December 21; the story was "There once was a strange little bug.")

– even though they may have rejected many of the individual choices. Figure 3.3 makes this point graphically, in terms of the average number of stories per student per day of Reading Tutor use. In fact, the number of stories finished per day reflected the desired 50/50 division of labor between the student and the Reading Tutor, *despite* students' rejection of initial Reading Tutor choices. On average, students finished 1.56 stories per day that were chosen by the Reading Tutor – and finished 1.54 stories per day that were chosen by the student (Figure 3.3).

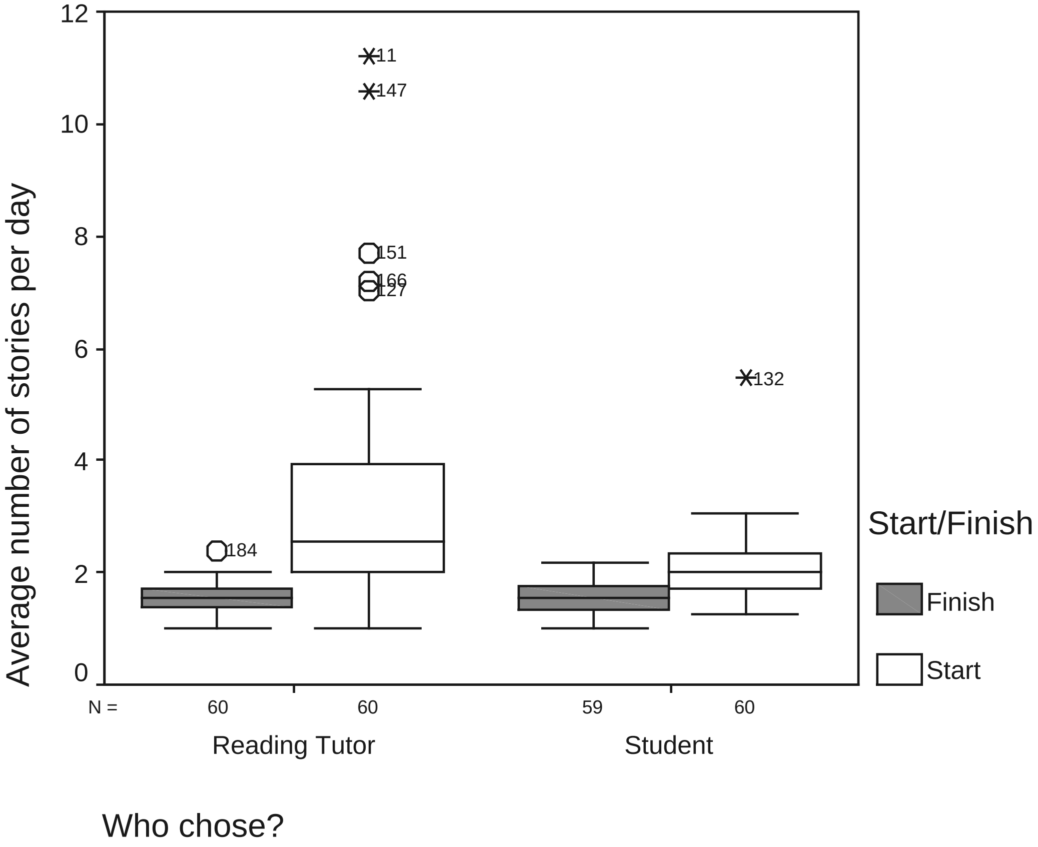


Figure 3.3. Average number of stories started and finished per day, by who chose the story.

3.3.3.2 Writing stories

In the 1999-2000 Reading Tutor, students could choose to write stories as well (Mostow &

Aist AAAI 1999, Mostow & Aist USPTO 1999). How large of a role did activities other than Reading Tutor stories play in students' story choices? Other activities were lumped together into level U: reading other students' stories, as well as writing and narrating the student's own stories. Therefore we examined the distribution of story levels. Figure 3.4 shows the breakdown of started and finished stories by student, who chose the story, and story level.

What is striking about the distribution of level U stories is how disparate the students' choices were. Many students chose level U stories rarely; 14 in fact did so not at all. On the other hand, 14 students chose level U stories as half or more of their story choices.

In order to characterize how often students were writing a story versus reading another student's story, we categorized each level U choice as either a student reading a different student's story or as writing a story. We categorized a level U choice as a student reading a different student's story if the first name, last initial, and inferred gender of the author as shown in the story title (e.g. Jane D.) did not match the recorded gender and initials of the student who selected the story. For example, Jane D. matches a female with initials J.D., but not a male with initials J.R. Otherwise, we categorized the choice as writing a story. Of the 996 choices of level U stories, 107 out of 996 choices were examples of students reading other students' stories. (Or, to look only at the first time a student selected a story with a particular title, 69 out of 242 distinct student-story pairs.) Appendix B gives further details on student-written stories. While interesting from other perspectives, what happened when students chose to write stories or to read other students' stories is not a key component of what we were trying to get the Reading Tutor to do. We simply remark that writing stories doesn't expose a student to new words, although reading some other student's might (in principle); we omit further discussion.

3.3.3.3 Story levels

What were the relative levels of what the Reading Tutor chose vs. what students chose? The stories that the Reading Tutor chose were generally harder than the stories that students chose. Consider stories that students actually finished: The per-student average grade level for Reading Tutor-chosen stories was 1.13 ± 0.65, where level K = 0 and level E = 6. The per-student average grade level for student-chosen stories was substantially lower: only 0.61 ± 0.59, $p < .001$ by T-test pairing each student's new-material percentage for Reading Tutor-chosen stories against his or her new-material percentage for student-chosen stories. The fact that many students were often able to finish stories that were harder than the ones they chose for themselves suggests that they were not challenging themselves enough. Alternately, students may have been able to get through stories chosen by the Reading Tutor even if the stories were too hard for the students to benefit from as much as they would have from easier stories. We also compared new student-chosen stories to previously read student-chosen stories. Looking at stories that students finished, we found that when students re-read stories, they re-read stories that were on average *lower* in level than the new stories they selected: 0.46 ± 0.66 for old stories, and 0.78 ± 0.57 for new stories, significant at $p < .001$ by T-test pairing data student-by-student. This result suggests that students preferred to re-read the easier stories.

Figure 3.4 shows the complete distribution of story choices, averaged on a per-student basis. Students might start a story and then not finish it for several reasons, such as: clicking *Back* to choose another story (or have the Reading Tutor choose another), clicking *Goodbye* to log off, or being automatically logged out because they were inactive for a long period of time. The high percentage of Reading Tutor choices that were not finished is most likely due to students clicking *Back* to reject the Reading Tutor's story choice.

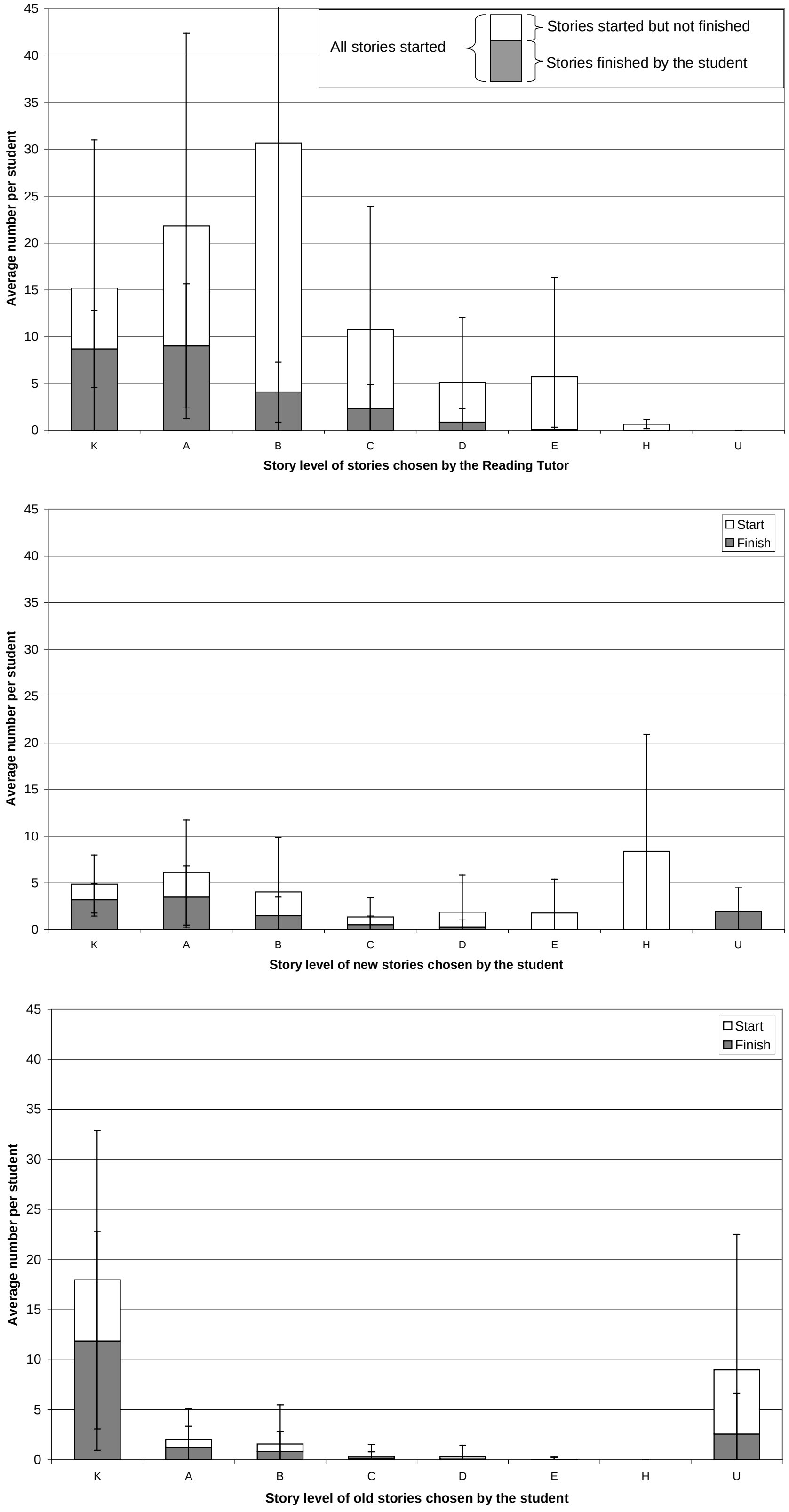


Figure 3.4. How many new stories were chosen and finished on average.

Figure 3.4 shows how many new stories were chosen and finished on average, by whether the student or the Reading Tutor picked the story, for Fall 1999. Level U story counts are based on considering stories the same if they have the same first line – thus potentially underestimating the number of new stories, if stories shared the first – Reading Tutor-supplied – line. Note that more than half of the new stories that were finished resulted from Reading Tutor choices, even though many of the Reading Tutor's initial choices were rejected.

3.3.3.4 Re-reading stories

How much re-reading was due to students re-reading a story once or twice, and how much was due to re-reading a favorite story? Figure 3.5 shows a bar chart of the number of times a story was re-read, and Figure 3.6 shows a closeup of the right side of Figure 3.5. These figures *exclude* level U stories, to look only at Reading Tutor stories that students read. The various patterns in each little box indicate different students. The boxes are too small to be fully legible, but the main point is that many students read a few stories a few times (Figure 3.5), but only a few students read one story many times (Figure 3.6). Those who did re-read a favorite story did so many times, however: as many as 17, 20, 28, and for one student even 32 times.

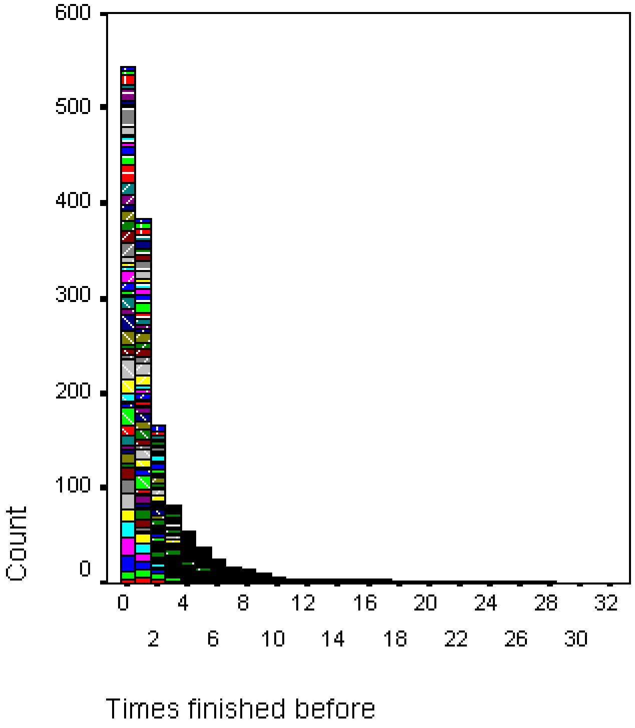


Figure 3.5. Graph of the number of times a student finished a story. The height of each little box is the number of stories that a student re-read a story *n* times, where *n* is times finished before as shown on the x axis. See Figure 3.6 for a closeup of the right hand side of the graph.

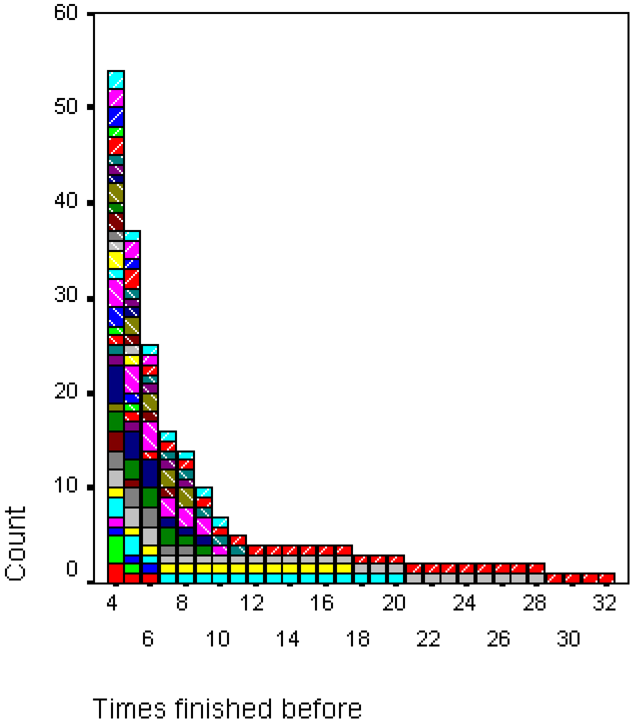


Figure 3.6. Closeup of Figure 3.5. showing story readings for students who finished the same story 4 or more times.

3.3.4 Efficiency

In previously published work (Aist & Mostow ITS-PA 2000) which we summarize here, we reported a comparison of the 1999-2000 Take Turns Reading Tutor to the student-only story choice Reading Tutor of Spring 1998. As we explained in Chapter 2, Section 1, the 1997-98 Reading Tutor required two separate clicks to select a story: one click to select a story, and another to confirm the choice. The revised Take Turns Reading Tutor required only one click to pick a story, and allowed the student to click *Back* to pick another story.

We set out to check that the Take Turns version of the Reading Tutor was at least as efficient as the previous student-only story choice version. In order to do this, we first compared the two versions of the Reading Tutor as follows. We wanted to filter out browsing behavior (choosing a story, clicking on *Back*, choosing another story, …) from our analysis in order to characterize how long it took to settle on a story to read. Thus we looked at the time it took to choose a story from the last sentence of one story to the *second* sentence of the next story. We (hand-)analyzed a small random sample of story choices. The student-only story choice Reading Tutor of Spring 1998 yielded story choice taking approximately 2 minutes (2.0 ± 1.7 minutes, 10 examples). Students using the Take Turns Reading Tutor in Fall 1999 took approximately 30 seconds to pick stories (0.44 ± 0.27 minutes, 9 examples), a substantial improvement and significantly better by a two-sample T-test (p = .02, assuming unequal variances.) We believe that both one-click pick and Take Turns contributed to this improvement.

The above analysis is limited, being based on a small sample – and also conflates the time it took to select a story with the time it took for the student to accept a choice. An alternate way to think about efficiency is just to compare how long it took for students to select a story in the 1997-98 Reading Tutor against how long it took the Reading Tutor to select a story. Here the difference is not a few percent, or even a factor of two or four: there is just no contest. The 1999-2000 Reading Tutor sorted through all the stories and made its choice quickly – under ten seconds in a sample of actual use – not even enough time for the student to listen to all of the story choices on the screen. Furthermore, while students' reading speed and the speed of spoken language will *always* limit how many story choices they can consider, computerized choices are not subject to any meaningful lower bound on how long the choice will take. (We are not talking about picking material from the World Wide Web with its billions of pages; we are talking about selecting from hundreds of stories.) Improvements in computer speed, algorithms, and pre-computing could make the computer's story choice even faster – whereas student story choice will always be limited to the speed of the student. Therefore the question of how fast an actual Reading Tutor story choice took is not really an issue for consideration when extrapolating to future usage.

While how much time an individual computerized choice takes is not of much concern, the total time that it takes to settle on a story choice is of some concern. In the Fall 1999 Reading Tutor, some factors may have led to increased time to settle on a story. For example, some children spent a lot of time browsing the story menu when it was their turn to pick – or rejecting many Reading Tutor choices when it was the Reading Tutor’s turn. Possible solutions include providing students with more information about the content of the stories they can choose from, improving the Reading Tutor’s choices to make them more palatable, or limiting the number of times that a student can turn down the Reading Tutor’s choice by clicking *Back*.

What ultimately matters, however, is how much reading students actually did because of the story choices they and the Reading Tutor made, as evidence that they were not simply spending all their time choosing stories. We therefore turn our attention to the keystone of the matter: effectiveness.

3.3.5 Effectiveness

In order to confirm that Take Turns got students to see relatively more material than the previous policy, we compared story choice data from the 1999-2000 Take Turns Reading Tutor versus the story choice data from the 1997-1998 Reading Tutor. We selected data from two approximately equal periods, with similar study conditions, as follows. We chose data from the spring of 1998: 24 students in grades 2, 4, and 5 at Fort Pitt Elementary School who used a version of the Reading Tutor with student-only story choice. For the 1999-2000 data, we chose data from the fall of 1999: 60 students in grades 2 and 3 at Centennial Elementary School who used the Take Turns Reading Tutor that took turns picking stories. We selected spring 1998 data and fall 1999 data in order to compare approximately equal groups:

- both groups of students spent 3-4 months with the Reading Tutor by the end of the period we examined;[2]
- in each study, there was one Reading Tutor per classroom, shared by 8-12 students;
- in both studies, students used the Reading Tutor individually in the classroom;
- all of the classrooms in each study were new to the Reading Tutor at the outset. (It is possible that some of the students in the Spring 1998 study had used the Reading Tutor the previous summer.)

Appendix A summarizes this experiment according to the coding scheme used by the National Reading Panel report *Teaching Children to Read* (NRP 2000).

How can we compare the effectiveness of the two versions? The stories available to students were different in 1997-98 vs. 1999-2000, and comparing the number of new words would reflect the changes in materials. Rather, we looked at the percentage of new sentences out of all the sentences seen – a measure of how much new material students were reading. This measure directly reflected the effect of story choice policy.

In order to compare the effectiveness of the two versions, we calculated the percent of new sentences that students saw, out of all sentences they saw. (The percentages were calculated on a per-student basis, to weight all students equally instead of favoring those who read more sentences.) Figure 3.7 shows boxplots for the percent of new sentences for the students in the Spring 1998 and Fall 1999 studies. In the remainder of this chapter, statistics on sentences were

[2] At Centennial Elementary, students continued beyond the Fall 1999 period we considered and used the Reading Tutor during the rest of the school year as well.

calculated by counting the files that recorded students' utterances, and include sentences from level U stories, which gives the benefit of the doubt to the value of student-authored U stories for vocabulary development. (The utterance files did not record the story title or level, so excluding sentences based on story level would be problematic.) Statistics on stories were calculated from the Reading Tutor's database, and exclude level U stories.

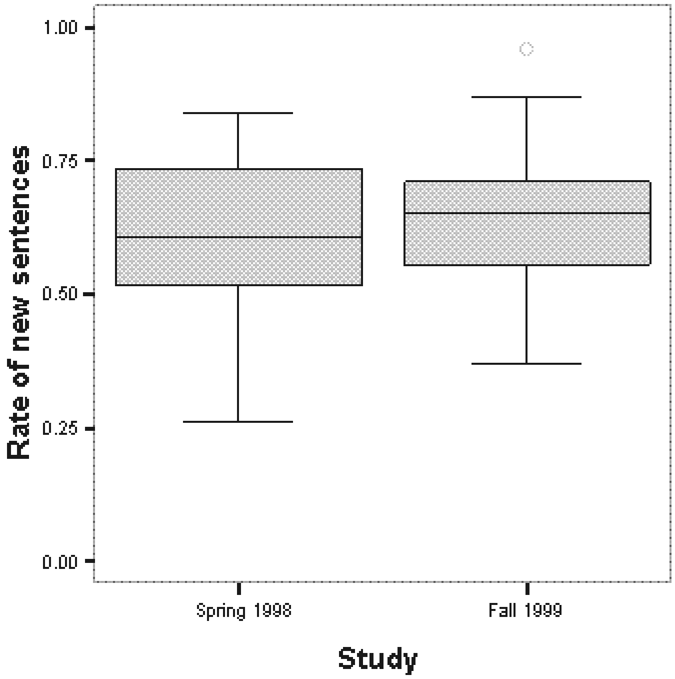


Figure 3.7. Boxplot[3] for per-student rate of new sentences seen out of all sentences, Spring 1998 and Fall 1999 studies.

The two studies included students in different (partially overlapping) grades: grades 2, 4, and 5 in Spring 1998 and grades 2 and 3 in Fall 1999. A univariate analysis of variance (ANOVA) with grade as a covariate revealed a significant difference in favor of Fall 1999 on the rate of new

[3] See glossary.

material encountered between the two conditions, significant at 90%: F=3.25, p = .075. We calculated the per-student average rate of encountering new sentences in each study, determined by first calculating each student's rate of encountering new material, and then averaging all the rates together. (Simply dividing the number of new sentences by the total number of sentences would be statistically incorrect because of a bias towards students who read more material.) In the spring of 1998, out of more than 10,000 sentences overall encountered by 24 students, the per-student average rate of seeing new sentences was 60.1%. In the fall of 1999, out of nearly 35,000 sentences encountered by 60 students, the per-student average rate of seeing new sentences was now 64.1%, a relative increase of 6.7%. Figure 3.8 shows boxplots of the rate of new material for the Spring 1998 and Fall 1999 studies, stratified by grade.

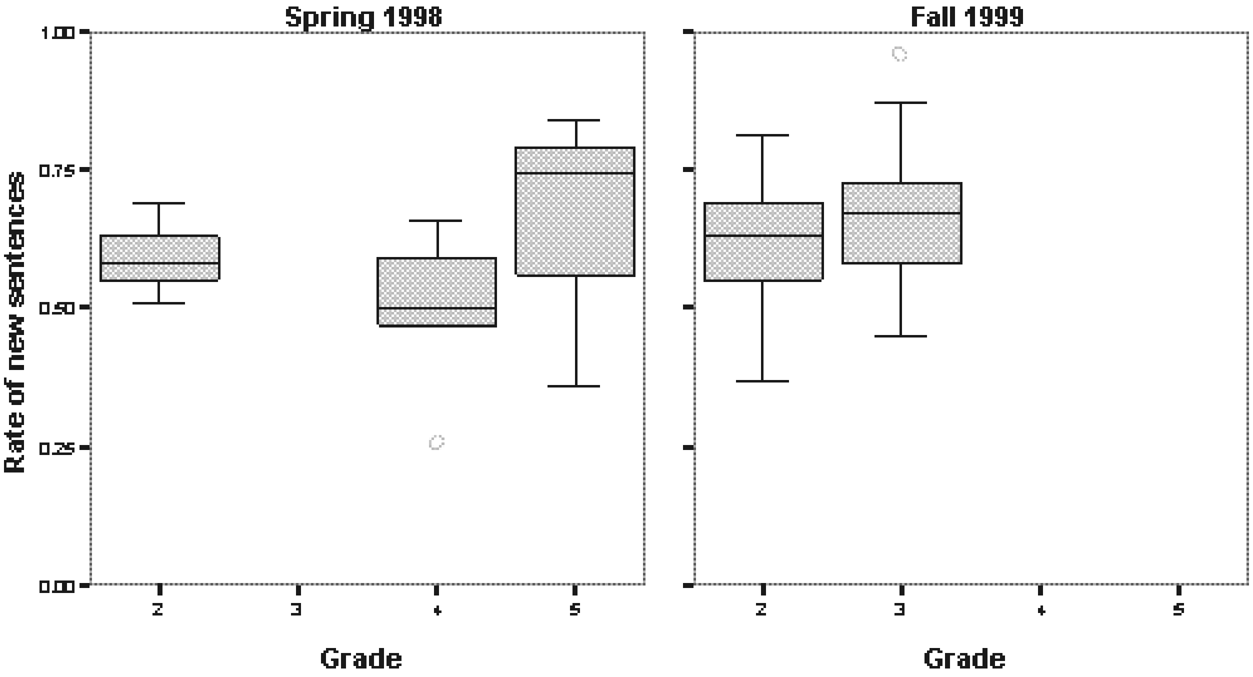


Figure 3.8. Spring 1998 student-only story choice vs. Fall 1999 Take Turns story choice for percentage of new sentences, subdivided by grade.

Further analysis revealed that the effect of the Take Turns policy was – as desired – to increase the amount of new material read by the students who read the least new material. Bottom-half students in the Fall 1999 "Take Turns" group showed a (very weak) tendency to read more new material than their peers in the Spring 1998 student-only story choice group (at p=.301 including grade as a covariate), while the top-half students in both groups read on average the same percentage (72%) of new sentences. In the Spring 1998 study, the 12 students below the median rate of seeing new sentences averaged 49.6% ± 9.7% new sentences, and the 12 students above the median averaged 72.1% ± 7.2% new sentences. In the Fall 1999 study, the 28 students below the median averaged 54.9% ± 6.7% new sentences, and the 32 students at or above the median averaged 72.2% ± 6.8% new sentences. Figure 3.9 makes this comparison graphically.

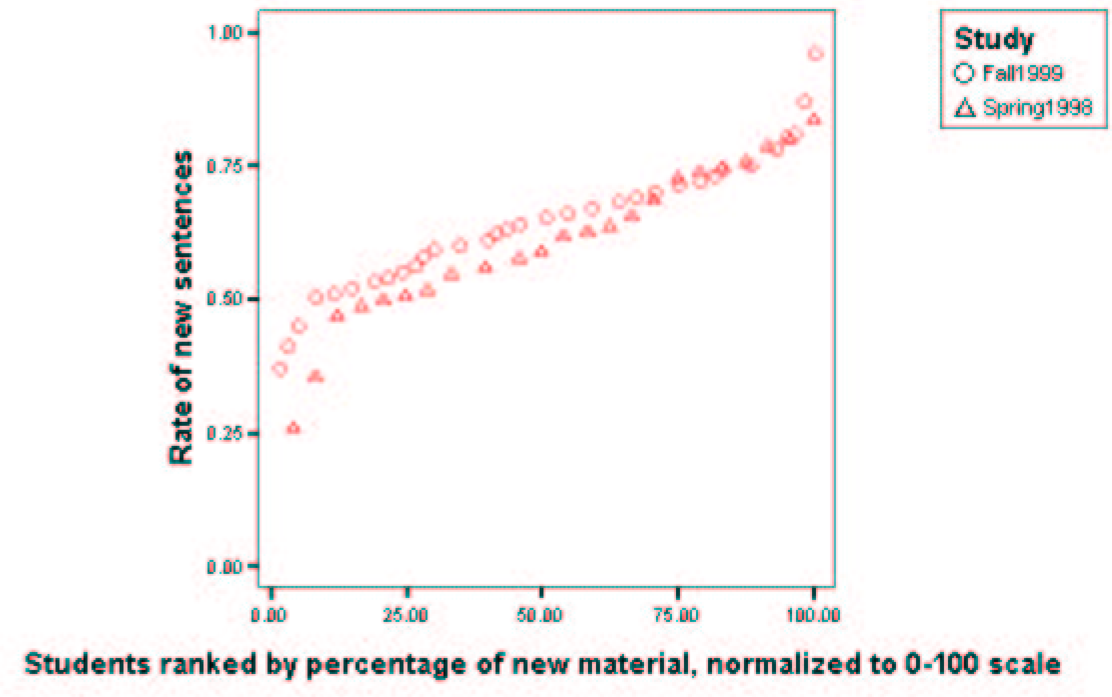
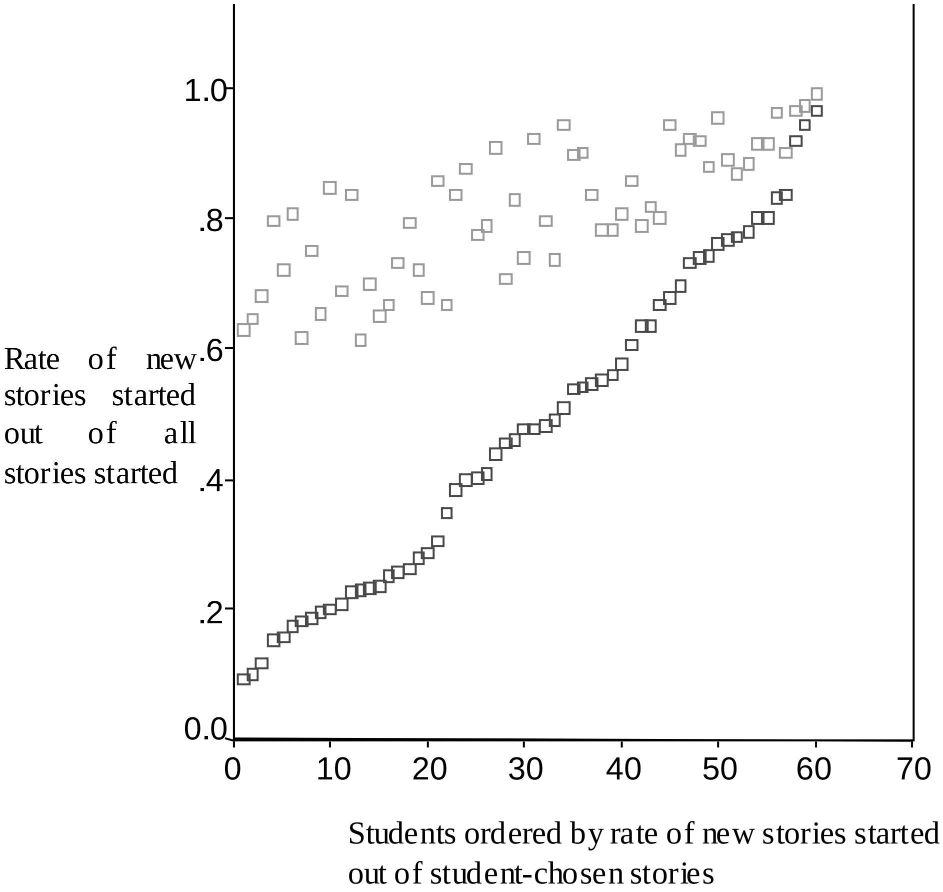


Figure 3.9 Spring 1998 vs. Fall 1999 study showing students ranked by percentile placement on the percentage of new material (new sentences / all sentences) they saw. (Some points may represent more than one student, due to ties in rankings.)

Finally, in order to check for the effects of the Reading Tutor's story choices on how many new stories students in Fall 1999 read, we looked at the Fall 1999 data in more detail. We looked at how often students chose new stories when it was their turn to pick vs. how often either the Reading Tutor or the student chose new stories (Figure 3.10). We also looked at the new-material percentage for student-chosen stories and overall (Figure 3.11).

For chosen stories and for finished stories, the added boost from Reading Tutor-chosen stories is impressive. For the stories chosen by students in the Fall 1999 study, fully half of the students had a new-material percentage of 43% for stories they finished. The bottom 10% of readers had on average a new-material percentage of less than 15% for stories they finished. When including stories chosen by the Reading Tutor, virtually all of the students read more than 50% new stories. Thus Take Turns helped the most for the students who chose the lowest percentage of new stories on their own.



.

Figure 3.10. New-material percentage for started stories, comparing stories chosen by the student (bottom points) vs. stories chosen overall (top points), Fall 1999.

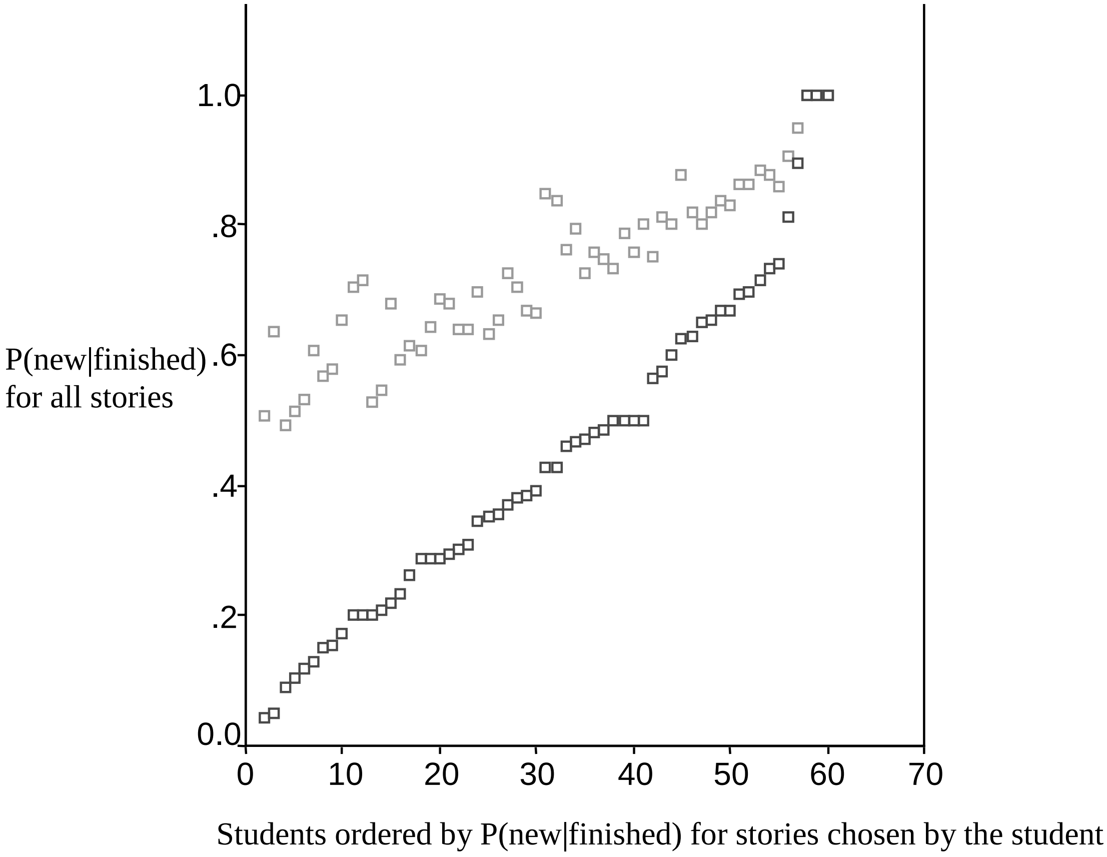


Figure 3.11. New-material percentage for finished stories, comparing student-chosen stories (bottom points) vs. percent new stories finished overall (top points), Fall 1999.

So students saw a higher percentage of new stories in the Take Turns version of the Reading Tutor, compared to the previous student-only story choice version. What about new words? Comparing new words directly across years has the problem that the materials changed between words – confounding the comparison. Nonetheless we would like to characterize how many new words students saw under the Take Turns policy. So how many words did students see in 1999-2000? (We look at the entire year because this figure bears on Chapter 5, where we will compare the Reading Tutor versus classroom instruction and versus one-on-one human tutoring.) Figure 3.12 shows boxplots of the number of distinct words seen by students in each grade. In second grade, the number of distinct words ranged from 458 to 1588, with a median of 1124. In third grade, the number of distinct words ranged from 856 to 1910, with a median of 1224. (Students thus saw about 11-12 new words per day, given that there were ~100 days in the 1999-2000 study.) Third graders saw more new words on average than did second graders, by analysis of variance (ANOVA) using number of new words and examining main effect of grade: (F=6.54, p=0.01). However, there was no significant difference for new stories read between second graders and third graders (F=0.81, p=0.37, Figure 3.13). What explains the difference? Second graders tended to read easier stories than third graders (Table 3.5), and easier stories tend to be shorter and yield fewer unique words.

	Reading Tutor	Student (new stories)	Student (old stories)
Second grade	Start: 1.65 ± s.e. 0.13 Finish: 0.89 ± s.e. 0.07	Start: 1.03 ± s.e. 0.16 Finish: 0.49 ± s.e. 0.09	Start: 0.21 ± s.e. 0.08 Finish: 0.18 ± s.e. 0.07
Third grade	Start: 1.96 ± s.e. 0.18 Finish: 1.39 ± s.e. 0.14	Start: 1.73 ± s.e. 0.13 Finish: 1.13 ± s.e. 0.09	Start: 0.92 ± s.e. 0.17 Finish: 0.75 ± s.e. 0.15

Table 3.5. Story level (K=0, A=1, …) by grade. All main effects significant at p < 0.001: grade (F=52.7, df=1), start/finish (F=30.0, df=1), who chose the story (F=17.8, df = 1), old vs. new story (F=42.0, df=1).

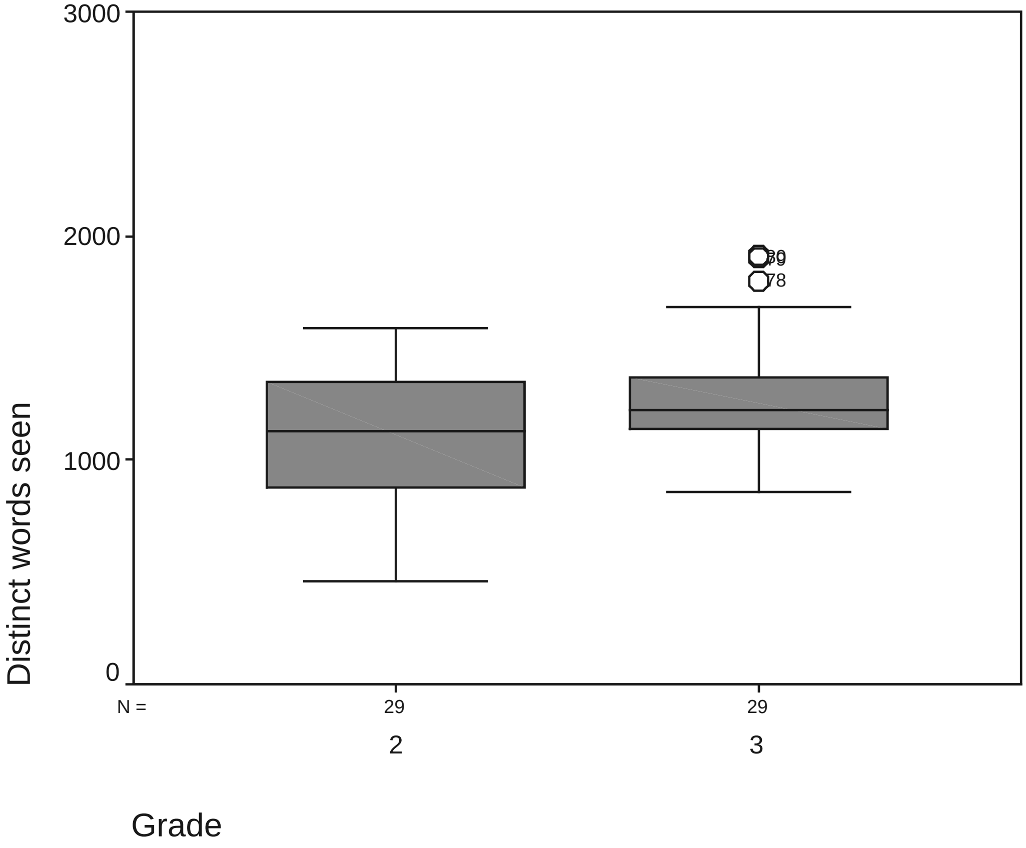


Figure 3.12 Number of distinct words (not stems) seen by students using the Reading Tutor in 1999-2000 study.

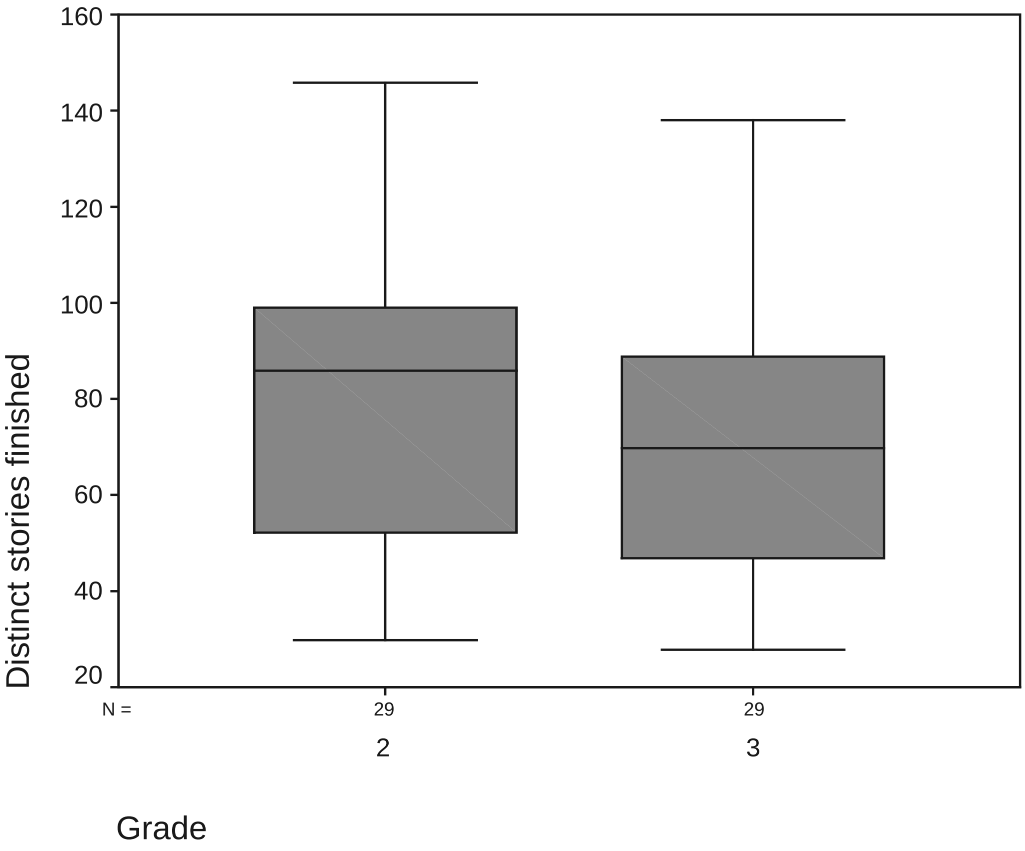


Figure 3.13 Distinct stories seen by students in 1999-2000 study.

3.4 Lessons learned from story choice experiments

One larger lesson we drew from the story choice experiment is that the nature of the data collected affected what comparisons were feasible to make. For a positive example, since the Reading Tutor collected the students' utterances, we were able to compare utterances across years – despite the various software changes between 1997-1998 and 1999-2000. For a negative example, it might be interesting and informative to compare the amount of time spent in various activities. But the Reading Tutor recorded a stream of single events, such as opening a story or finishing a story, not intervals – which require additional bookkeeping and may not always have well-defined start and endpoints. Uncertainty about when an opened story ended in the event of a timeout, program restart, computer crash, or other situation make interval determination less straightforward and more prone to error than comparing percentages of recorded events. Thus we have used percentages calculated directly over events – the data representation supported most directly by the raw, field-collected data – to capture the essence of Reading Tutor use. Furthermore, we supplemented automated analysis in certain cases with small, hand-counted samples in order to obtain a complementary, common-sense view of the experimental results.

In conclusion, Take Turns was demonstrably better than student-only story choice at ensuring that all students read new material, and at increasing the amount of new material that students read. Take Turns helped students read new stories, especially those students who read the least new material on their own.

However, as we pointed out in Chapter 1, making sure that students encounter new words doesn’t guarantee they’ll learn the meaning of the new words from the encounters. How can we help students learn more from encounters with words? In the next chapter, we address the question of giving vocabulary assistance.

4 Automatically generated vocabulary assistance and assessment: The factoids experiment

Reading material that contains new words is a requirement for learning new words from reading text. However, simply reading new and challenging stories may not be sufficient. Individual encounters with a word may not contain enough information to learn much about the word. How can text be augmented so students can learn more from an encounter with a word than they would have otherwise – yet without taking too much additional time away from reading the original text?

We decided to explore augmenting text with vocabulary assistance. In the experiment described in this chapter, we compared augmented text to unaugmented text, rather than to a "no exposure" control – because if the augmentation does not help over and above unaugmented text, adding augmentation would probably just waste the student's time.

We now discuss several design questions for giving vocabulary help, and discuss where our vocabulary experiments stand in relation to each of them.

Which students need vocabulary help? Identifying which students could benefit from vocabulary help turned out to be one of the issues our experiments shed light on. In our experiments, we gave vocabulary help to elementary students in various grades and at different reading levels.

For which words should the Reading Tutor give vocabulary help? In our studies we explored various categories: words with few senses (here in Chapter 4), words important to the meaning of a passage (in Chapter 6, Section 1), and domain-independent – but very rare – words (in Chapter 6, Section 2).

What kind of vocabulary help should the Reading Tutor give?

Options include:

- A conventional definition. “*as·tro·naut*. A person trained to pilot, navigate, or otherwise participate in the flight of a spacecraft” (American Heritage 3rd edition, 1996).
- A definition from a children’s dictionary. Definitions may vary widely in length and difficulty. For example, this definition for *astronaut* is short and sweet: “astronaut. a traveler in a spacecraft” (Merriam-Webster Student Dictionary, wordcentral.com). But consider the definitions for *comet* and *meteor*: “comet. a bright heavenly body that develops a cloudy tail as it moves in an orbit around the sun”; “meteor. one of the small bodies of matter in the solar system observable when it falls into the earth’s atmosphere where the heat of friction may cause it to glow brightly for a short time; also: the streak of light produced by the passage of a meteor” (Merriam-Webster Student Dictionary, wordcentral.com).
- A comparison to another word. “An *astronaut* is a kind of traveler.”
- A short explanation. “An *astronaut* is someone who goes into outer space.”
- An example sentence. “The *astronaut* went to the Moon in a rocket.” See for example Scott and Nagy (1997).

In this chapter, we describe work on comparisons to other words. (Chapter 6 describes experiments on short explanations.)

At what time should help be given? When should the Reading Tutor provide vocabulary help on a word in a story – before the student reads the story, during the story, or after the story? For high school readers, Memory (1990) suggests that the time of instruction (before, during, or after the reading passage) for teaching technical vocabulary may not matter. If the lack of difference between presentation times holds true for elementary students as well, the Reading Tutor may be able to choose from several different times to give vocabulary help, without diminishing the student's ability to learn from the assistance. In the study described in this chapter, we inserted vocabulary assistance just before the sentence containing the target word. (One study described later, in Chapter 6, involved inserting vocabulary assistance before the story (limerick) containing the target word.)

Who should decide when help is necessary – the computer, the student, or both? We wanted to focus on the effects of vocabulary assistance unconfounded by whether the student requested help. Thus, in order to provide help equally to students who click frequently and to those who rarely click at all, we further chose to have the computer (or designer) control the presentation of words, rather than display explanations at the student's request.

In the remainder of the chapter we describe an experiment on vocabulary assistance: automatically generating and presenting comparisons to other words, and automatically generating and administering assessments. We discuss rationale, design, implementation, results, and lessons learned.

4.1 Rationale

We wanted to add vocabulary assistance to text to make computer-assisted oral reading more effective for word learning. We did not intend to *replace* reading text with studying synonyms, as some previous studies have done (Gipe & Arnold 1978). Instead, we *augmented* assisted reading with comparisons to other words the student might already know. By analogy, consider salt: salt augments flavor, so salt is added to food – not used instead of food. Likewise, we did not contemplate completely replacing assisted reading with practice on synonyms – just augmenting text with semantic information to give students a learning boost when they encountered novel words.

4.1.1 Automatic generation of vocabulary assistance

We wanted to make vocabulary assistance that was applicable to any text. To do so, we needed a large-scale resource to cover many words students would encounter over the course of months of Reading Tutor use. We needed both to provide assistance and to assess its effects. To meet the goal of large-scale assistance and assessment applicable to any English text, we made use of a well-known lexical database: WordNet (Fellbaum et al. 1998). WordNet, originally developed by George Miller and colleagues, contains tens of thousands of words organized by a thesaurus-style hierarchy (*astronaut* is a kind of *traveler*) and with links to synonyms (*astronaut* and *cosmonaut* are synonyms in WordNet). We designed automated assistance, applicable to any text, that compared words in the text to other words in WordNet.

Such comparisons make most sense where the new word is a specialized form of some previously known concept, such as *morose* meaning very sad. They might not be expected to work as well on vocabulary describing previously unknown concepts, except to the extent that kind-of relations apply, such as *tuberculosis* being a kind of disease.

4.1.2 Automatic generation of vocabulary assessment

We also needed to evaluate the effectiveness of vocabulary assistance. Nagy, Herman, and Anderson (1985) categorize multiple-choice questions according to how close the distractors (incorrect answers) are to the correct answer. Nagy, Herman, and Anderson’s classification is as follows:

Level 1. Distractors are a different part of speech from the correct answer. For example, if the target word is *astronaut* and the correct answer is *traveler*, Level 1 distractors might be *eating*, *ancient*, and *happily*.

Level 2. Distractors are the same part of speech but semantically quite different. For example, if the target word is *astronaut* and the correct answer is *traveler*, Level 2 distractors might be *antelope*, *mansion*, and *certainty*.

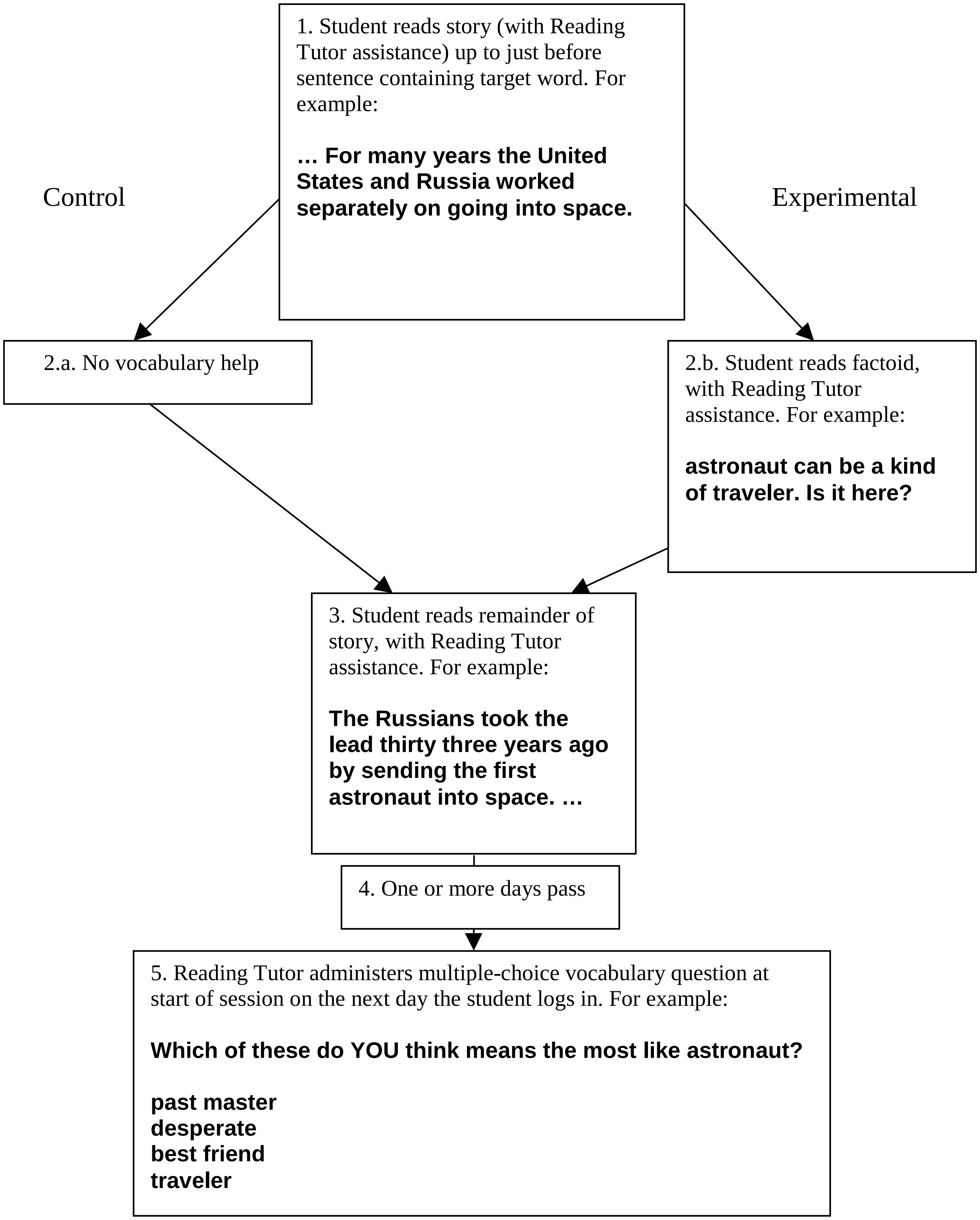
Level 3. Distractors are semantically similar to the correct answer. For example, if the target word is *astronaut* and the correct answer is *traveler*, Level 3 distractors might be *doctor*, *lawyer*, and *president*. This example illustrates that sometimes it is easier to design the intended answer and the distractors together; a more natural example would be a correct answer of (*space*) *pilot*, with the same distractors *doctor*, *lawyer*, and *president*.

We designed automated vocabulary assessment questions using the WordNet hierarchy, taking as our goal Nagy, Herman, and Anderson's Level 3 multiple choice questions. Section 4.3.1 provides further details.

Skeptics might ask: Why use automated, variable-quality, experimenter-designed questions instead of a standardized instrument? First, our decision to use experimenter-designed questions was subsequently validated by the National Reading Panel's (later) call for the use of experiment-defined measures to test vocabulary. Experimenter-constructed measures, which often measure gains on particular words, tend to be more sensitive to small gains than standardized tests that aim to measure vocabulary in general by sampling a small number of words (NRP 2000). Second, we used this measure as a comparison inside the interaction. That is, all students received assistance, saw words in the experimental and control conditions, and took the multiple-choice tests. Thus our experiments were within-subject in a way that standardized tests do not (easily) facilitate. Third, we did in fact pre-test the students using a widely used measure of vocabulary – the Word Comprehension subtest of the Woodcock Reading Mastery Test (WRMT, see chapter 5 for details). As a test of external validity, we calculated the correlation between students' performance on the multiple-choice questions for words seen in context (without extra assistance), and their (grade-normed) performance on the Word Comprehension section of the WRMT. The correlation was significant, at r=0.47 for grade 2 (p = .009) and r=0.49 for grade 3 (p = .008). Thus our measure fit the national research agenda, could be automatically constructed and scored, and was correlated with a widely used external test of vocabulary knowledge.

4.2 Experiment design

Figure 4.1 shows the design of the experiment described in this section, intended to contrast seeing a word in a story alone vs. seeing a word in a story along with some vocabulary help.



1. Student reads story (with Reading Tutor assistance) up to just before sentence containing target word. For example:

… For many years the United States and Russia worked separately on going into space.

2.a. No vocabulary help

2.b. Student reads factoid, with Reading Tutor assistance. For example:

astronaut can be a kind of traveler. Is it here?

3. Student reads remainder of story, with Reading Tutor assistance. For example:

The Russians took the lead thirty three years ago by sending the first astronaut into space. …

4. One or more days pass

5. Reading Tutor administers multiple-choice vocabulary question at start of session on the next day the student logs in. For example:

Which of these do YOU think means the most like astronaut?

past master
desperate
best friend
traveler

Figure 4.1. Factoid flowchart, showing one example using the target word *astronaut*. (*Desperate*, like the other three possible answers, is used as a noun here.)

In summary, a control trial was as follows:

1. First, the student read out loud (with the Reading Tutor's assistance, as described in Chapter 2) the portion of the story up to but not including the sentence containing the target word.
2. a. Second – this is the control condition – nothing happened.
3. Third, the student continued reading the story, starting at the sentence containing the target word.
4. Fourth, one or more days elapsed.
5. Finally, the student answered a multiple choice question on the target word.

An experimental trial was as follows:

1. First, the student read out loud (with the Reading Tutor's assistance) the portion of the story up to but not including the sentence containing the target word.
2. b. Second – this is the experimental condition – the student read out loud (with the Reading Tutor's assistance) a factoid comparing the target word to another word. For example, "astronaut can be a kind of traveler. Is it here?"
3. Third, the student continued reading the story, starting at the sentence containing the target word.
4. Fourth, one or more days elapsed.
5. Finally, the student answered a multiple choice question on the target word.

Here is an example of an experimental trial, excerpted from actual Reading Tutor use during Fall 1999. For convenience, we have put events involving the target word *astronaut* in boldface.

	Time event occurred	What happened?
	Wednesday, October 6, 1999 12:37:10.356	Student (P.O., girl aged 9 years 5 months) chooses Level C story "Life in Space" (adapted from a Weekly Reader passage)
2 seconds later	12:37:12.259	Reading Tutor displays sentence "For many years the United States and Russia worked separately on going into space." Student tries reading sentence out loud.
19 seconds later	12:37:31.106	Student finishes speaking. Actual utterance:

		for many years the united states of russia worked s... sponidy on going to space Reading Tutor heard: FOR MANY YEARS THE UNITED STATES AND RUSSIA WORKED SEPARATELY SEPARATELY ON GOING INTO SPACE (The Reading Tutor's hearing is not perfect; in this case, it may have not detected the miscue because "sponidy" sounded more like "separately" than like the other words in the sentence (or truncations thereof), which is all the Reading Tutor listened for.)
< 1 second later	12:37:31.166	Reading Tutor decides to display next sentence of story
24 seconds later	**12:37:55.391**	**Reading Tutor displays first sentence of factoid: "astronaut can be a kind of traveler."** **Student tries reading sentence.**
16 seconds later	12:38:11.464	Student finishes speaking; Reading Tutor heard: ASTRONAUT CAN BE A KIND OF TRAVELER ASTRONAUT CAN BE A KIND OF TRAVELER
< 1 second later	12:38:11.524	Reading Tutor decides to go on to the next sentence
3 seconds later	12:38:14.408	Reading Tutor displays second sentence of factoid: "Is it here?"
9 seconds later	12:38:23.571	Student finishes speaking; Reading Tutor heard: IT INDIA IS IT HERE (What the Reading Tutor heard was not necessarily what the student actually said. If the sentence was short, the Reading Tutor included additional words to listen for, to approximate students' oral reading insertions and deletions, and to reduce acceptance of incorrect student attempts. Here, one "extra" word was INDIA.)
< 1 second later	12:38:23.621	Reading Tutor decides to display next sentence
1 second later	**12:38:24.843**	**Reading Tutor displays: "The Russians took the lead thirty three years ago by sending the first astronaut into space."**
		[Time passes]
Almost 24 hours later	Thursday, October 7, 1999 12:28:06.621	Student logs in the next day
2 seconds later	12:28:08.564	Reading Tutor presents student's name, for student to read as confirmation of identity and to make sure the microphone is working
10 seconds later	12:28:18.098	Student finishes reading name
9 seconds later	12:28:27.581	Reading Tutor presents vocabulary question by displaying the question and the answers, reading the question and then the answer out loud. Which of these do YOU think means the most like pail? railway car; paper bag; bucket; piles
16 seconds later	12:28:43.845	Student clicks on *bucket* (right!)
6 seconds later	12:28:49.713	Reading Tutor presents vocabulary question: Which of these do YOU think means the most like asparagus? butterfly pea; bog plant; yam plant; herb
20 seconds later	12:29:10.232	Student clicks on *herb* (right!)
17 seconds later	**12:29:36.881**	**Reading Tutor presents vocabulary question:** **Which of these do YOU think means the most like astronaut?** **past master; desperate; best friend; traveler**
17 seconds later	**12:29:54.025**	**Student clicks on *traveler* (right!)**
5 seconds later	12:29:59.013	Reading Tutor presents vocabulary question: Which of these do YOU think means the most like fetch? bring; project; impact; ferry
14 seconds later	12:30:13.073	Student clicks on *impact* (wrong!)
4 seconds later	12:30:17.299	Reading Tutor presents vocabulary question: Which of these do YOU think means the most like silk? material; hill; piece; cross

21 seconds later	12:30:37.708	Student clicks on *material* (right!)
8 seconds later	12:30:45.760	Reading Tutor chooses Level A story: “The Letter A”

A few notes on this example: First, displaying factoids sometimes caused delay due to database access. (In the case of *astronaut* in this example, 24 seconds). Second, it was not unusual for students to repeat a sentence if the Reading Tutor did not immediately accept their reading.

The Reading Tutor showed one factoid for every experimental trial. Thus, if two target words were in a single sentence, and both target words were randomly assigned to the experimental condition, the Reading Tutor would show a separate factoid for each target word. (If a target word appeared twice in the same sentence, we intended for the factoid to be displayed only once; we discovered after the fact that such factoids were not displayed at all – an unfortunate occurrence, but one that would have occurred only on those – presumably few – sentences where the same target word occurred twice.)

A minor bug caused the Reading Tutor to display multiple factoids for certain words, namely just those words which occurred as the first word of the sentence, capitalized. These few words (less than ten) were excluded from the analysis of the experiments.

We now discuss in more detail how the Fall 1999 Reading Tutor carried out the factoid experiment.

4.3 Implementation of the factoid experiment

We now summarize the factoid assistance.

Which students? Second and third graders using the Reading Tutor in Fall 1999 at Centennial Elementary School, near Pittsburgh, Pennsylvania. (Students actually used the Reading Tutor during the whole school year, but this experiment was only active during Fall 1999.)

Which words? The Reading Tutor selected as target words those words with three or fewer senses in WordNet, subject to some other constraints described in more detail below.

What kind of help? The Reading Tutor provided automatically generated comparisons to other words, using WordNet. For example, "astronaut can be a kind of traveler. Is it here?" The Reading Tutor presented a factoid as text for the student to read with the Reading Tutor's regular assistance as described in Chapter 2.

At what time to give help? The Reading Tutor presented the factoid just prior to the sentence containing the target word.

Who decides when help is needed? The Reading Tutor assigned words to control or experimental trials, randomized for each student.

In order to test the effects of this assistance, the Reading Tutor administered multiple choice questions on a later day. We now describe in detail how the Reading Tutor generated vocabulary assistance.

4.3.1 Selecting target words

A target word for factoid vocabulary assistance had to meet several conditions.

4.3.1.1 Conditions for generating vocabulary assistance

The Reading Tutor had to be able to give automated help on the word. Words with many senses might have resulted in too-long or confusing help. Thus, the word had to have only a few senses in WordNet: three (3) or fewer senses, including those of all parts of speech. Senses included those of the stemmed version (e.g. *time*) as well as from the actual text token (e.g. *times*). Stemming was done using WordNet's stemming function, called "morph". For a positive example: *astronaut* could be a target word because it has one sense: "astronaut, spaceman, cosmonaut -- (a person trained to travel in a spacecraft; 'the Russians called their astronauts cosmonauts')" (WordNet 1.6). For a negative example: *times* could not be a target word because while *times* has only the two senses "*multiplication*" and "*best of times, worst of times*", *time* has 14 senses).

The word could not be a trivially easy word. The word must have been three or more letters long – a heuristic designed to filter out trivially easy words such as *cat* or *dog*. The word could not be on a list of 36 words given by Mostow et al. (1994), shown in Table 4.1. In addition, the word must not have been a number written in Arabic numerals (for example, *200* or *35*.)

a	all	an	and	are	as	at	be	by	for	he	her
him	his	I	if	in	is	it	its	me	not	of	off
on	or	she	so	the	them	then	they	this	to	we	you

Table 4.1. Thirty-six function words excluded from vocabulary experiment.

The word could not be a proper noun. The word could not be a capitalized word (except for the first word in the sentence, which may be capitalized). This heuristic was designed to eliminate most names.

4.3.1.2 Conditions for generating vocabulary assessment

The Reading Tutor had to be able to ask a vocabulary question about the target word. To ask a multiple choice question, the Reading Tutor needed a correct answer and three distractors. We aimed at operationalizing Nagy et al.'s criterion of semantically similar distractors (Nagy, Herman, and Anderson 1985). To construct a 4-item multiple-choice vocabulary question, the Reading Tutor needed the correct answer and three wrong answers to serve as distractors.

The Reading Tutor used synonyms and hypernyms as the correct answer, reverting to a sibling (Figure 4.2) only if neither a synonym nor a hypernym could be found.

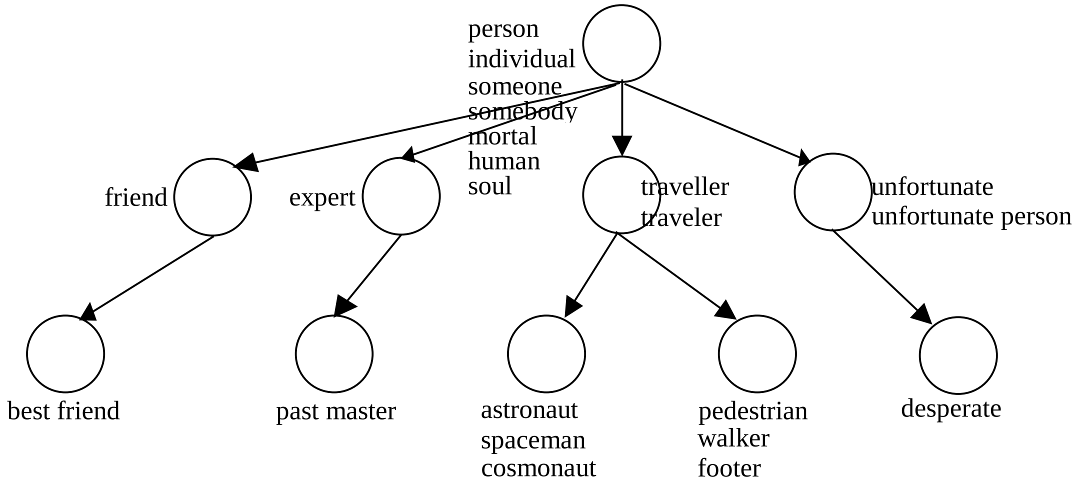


Figure 4.2. Siblings and cousins in WordNet. Selected portion of the WordNet 1.6 hierarchy. Nodes contain the set of all words that are synonyms of each other – that is, that form a single synset. Arrows point from general to more specific nodes. *astronaut* is a sibling to *pedestrian* because their nodes share the parent node "*traveller*; *traveler*". *astronaut* is a cousin to *best friend* because their nodes share the grandparent node "*person*; *individual; … soul*".

The word must have at least three vocabulary question distractors. Vocabulary question distractors were cousins of the target word (words with a common grandparent but different parents) (Figure 4.2.) The distractors were chosen so that the multiple choice question tested a student's ability to select the meaning of the target word from several semantically similar alternatives.

On the actual test, the answers were in a randomized order. (Random order was a potential source of variance, but reduced possible effects of children seeing their peers answering questions on the same words, when it was their turn to use the Reading Tutor.) Also, the selection of a particular correct answer and distractors was not constant for a word, but chosen anew for each trial.

4.3.1.3 Social acceptability

The word had to have been socially acceptable. The target word, the comparison word, the intended answer, and the distractors all had to be socially acceptable. We screened for acceptability in two ways. To forestall obviously offensive words, we required a natural-speech narration of a target word to have been recorded beforehand by a Project LISTEN team member, since we trusted project members not to record inappropriate words. To exclude words that are fine to pronounce, but risky to give automatically generated semantic help on (such as words with secondary slang meanings), the word must not have been on a list of explicitly banned words. These heuristics avoided the most egregious problems – but they were not perfect, allowing phrases where each word in itself was inoffensive, but the entire phrase was not: for example, *white trash* – a cultural slur.

4.3.2 Assigning words to conditions for factoid vocabulary assistance

We now describe how the Reading Tutor assigned words to conditions during the Fall 1999 factoid vocabulary study. For each student, half of the target words were randomly assigned to the experimental condition (factoid plus context), and the rest of the target words to a control condition (context alone). This randomization was done on a per-student basis. Thus while one student might see *astronaut* in the experimental condition, another student might see *astronaut* in the control condition. When the student encountered a previously unseen target word, the Reading Tutor assigned the new word to either the experimental (factoid plus context) or control (context alone) condition for that student. Since the same passages were used in control and experimental trials, this experiment controlled for text and word differences by randomly counterbalancing across trials, and relied on thousands of trials to wash out variance. While ultimately we might want to select words to explain based on what words are important to explain to which students, the Fall 1999 Reading Tutor used a blind, random assignment of words to conditions intended to persist for a given student's experience.

Treatment fidelity. By having an open-ended set of target words instead of a fixed list, we enabled the Reading Tutor to give assistance, without disrupting the study design, on any new material added by teachers, students, or the Project LISTEN team. The assignments of words to conditions were intended to persist throughout a particular student's history of Reading Tutor use to enable us to look for longer-term effects of multiple exposures to a word. Unfortunately, due to a flaw in the software, the assignments were not saved to disk. We therefore analyzed only a student's first day of experience with a word, and the subsequent vocabulary question.

4.3.3 Constructing vocabulary assistance

We now describe the form, content, and implementation of the factoid vocabulary assistance.

The Reading Tutor displayed vocabulary help for target words in the form of short comparisons to other words. The other words were extracted from WordNet. The vocabulary help was hedged because it might have been incorrect. For example, the comparison word might have been related to a different sense of the word than actually appeared in the story. The hedge question also aimed to encourage the student to think about the meaning of the word in context. For example: “astronaut can be a kind of traveler. Is it here?”

The Reading Tutor constructed the text of the factoid from a template containing placeholders for the target word and for the comparison word. The templates used in the 1999-2000 study were as follows.

Antonym. “The_Stem may be the opposite of The_Antonym. Is it here?”

Hypernym. “The_Stem can be a kind of The_Hypernym. Is it here?”

Synonym. “Maybe The_Stem is like The_Synonym here... Is it?”

Here, The_Stem was the base form of the word (*astronauts* → *astronaut*), The_Antonym was a word meaning the opposite of the target word, The_Hypernym was a more general word than the target word, and The_Synonym was a word that meant the same as the target word. Hypernyms and synonyms were used more frequently than antonyms. (Like many words, *astronaut* has no generally accepted opposite.)

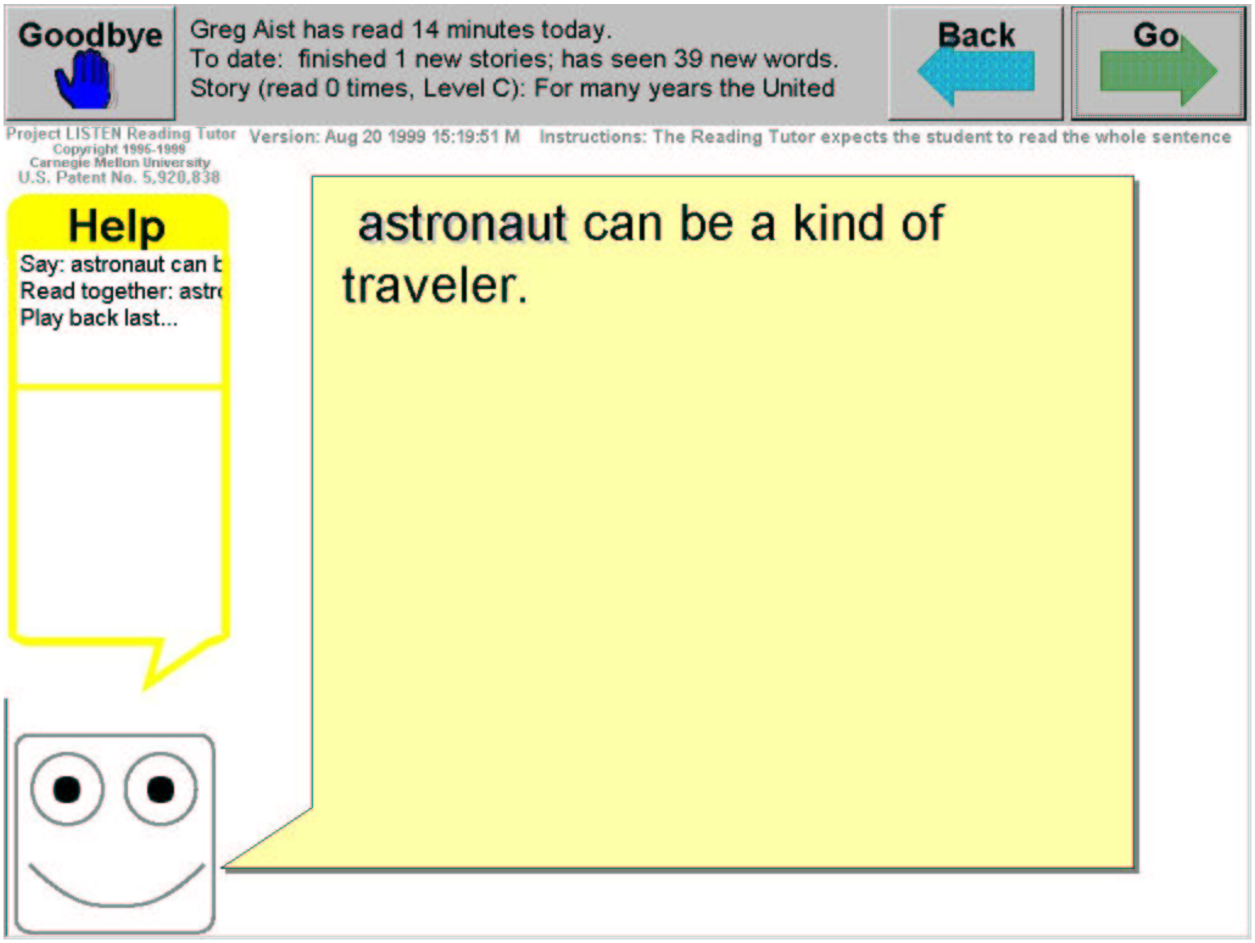
4.3.4 Presenting vocabulary assistance

To make sure that the student would pay attention to the vocabulary assistance, and to give the student extra practice in reading the target word, we presented the vocabulary assistance as text for the student to read out loud with the Reading Tutor’s help. (Other possibilities we considered included simply speaking the vocabulary assistance, presenting the text briefly in a drop-down window below the original sentence, or some combination of spoken and drop-down text.)

We also had a number of other design goals which were met in extended joint design work with the present author and Project LISTEN team members, especially Kerry Ishizaki and Jack

Mostow; also Human-Computer Interaction Masters' student Margaret McCormack.

- To distinguish the factoid from the original text, the factoid appears on a yellow background.
- To attribute the factoid to the Reading Tutor instead of the author of the original text, the factoid appears in a call-out balloon attached to the face in the lower left hand corner of the screen.
- To avoid confusion about what to read, and to simplify layout, the balloon occludes the original text.
- To provide first-class assistance, the factoid is presented as text for the student to read aloud, with Reading Tutor assistance (Figure 4.3). (Presenting the factoid as text to read also allowed for the possibility of giving factoids on factoids – we didn't, but might want to in the future.)



Greg Aist has read 14 minutes today.
To date: finished 1 new stories; has seen 39 new words.
Story (read 0 times, Level C): For many years the United

Project LISTEN Reading Tutor
Copyright 1996-1999
Carnegie Mellon University
U.S. Patent No. 5,920,838

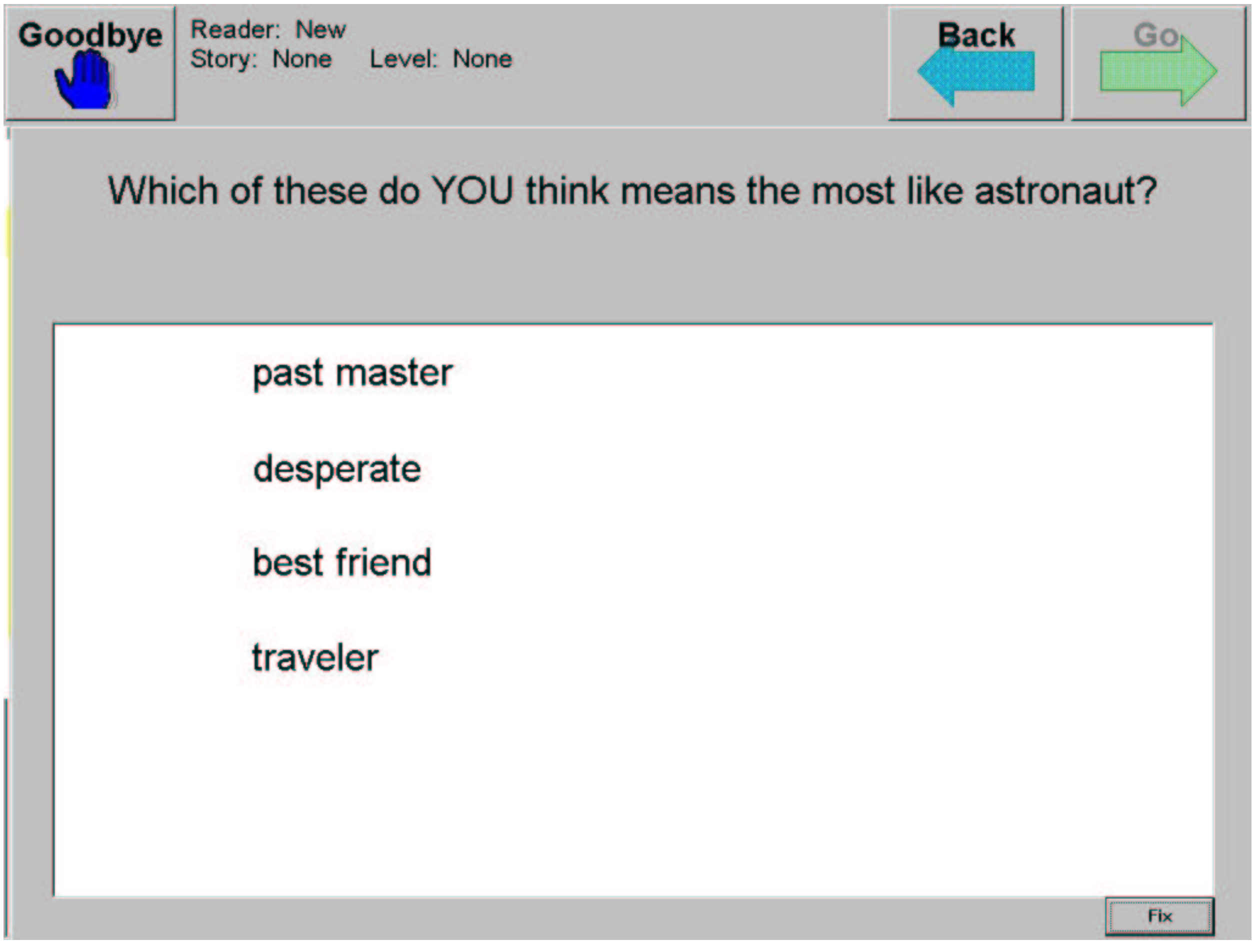
Version: Aug 20 1999 15:19:51 M Instructions: The Reading Tutor expects the student to read the whole sentence

astronaut can be a kind of traveler.

Figure 4.3. Factoid in popup window.

4.3.5 Administering the automatically constructed multiple-choice questions

We assessed the effectiveness of vocabulary intervention as follows. The next time a student logged in (one day or more after seeing the target word) the Reading Tutor asked vocabulary questions for each of the target words the student had encountered – both experimental and control words. Figure 4.4 shows a multiple-choice vocabulary question as displayed by the Reading Tutor. The target word was *astronaut*; the answers were: *past master, desperate, best friend, traveler*. (The correct response is *traveler*.) The answers were displayed in a random order, to prevent bias and hamper cheating, as explained earlier. The Reading Tutor spoke the prompt at the top of the screen, and then spoke the answers one at a time while highlighting each answer in yellow. The student could select an answer at any time by clicking on it; nonetheless the vocabulary questions did take time to answer (ranging from 14-21 seconds each in the examples above). Since the vocabulary questions were administered at the start of the session, they could not be contaminated by additional assistance just before being asked.



Reader: New
Story: None Level: None

Which of these do YOU think means the most like astronaut?

past master

desperate

best friend

traveler

Figure 4.4. Multiple-choice question for factoid experiment.

4.3.6 Data integrity

We used a database to collect the data from the over 3000 factoid trials. One trial was not properly recorded to the database due to a full hard drive: on Wednesday, March 22, 1999, a student received help on the word POUNCE that was recorded in the Reading Tutor’s log file, but not in its database.

4.4 Results of the factoid experiment

How much did factoids help? In order to assess the effectiveness of factoid assistance overall (3359 trials), we compared student performance on the experimental condition (factoid + context, 1679 trials) to student performance on the control condition (context alone, 1680 trials). Individual students' performance on all conditions ranged from 23% to 80%, with chance performance at 25% (1 out of 4).

The National Reading Panel Report remarked that many reading studies choose the wrong value of N (NRP 2000). In this case, analyzing this experiment using the trials as independent data points would be statistically incorrect, because a given student's trials were not independent of one another, and also because the number of trials varied from student to student. Analyzing this experiment by direct comparison of per-student averages would *underestimate* the effective sample size, because the average is not a single measure but rather a composite of multiple related trials. Logistic regression models offer a statistical technique for representing multiple responses from multiple students, and analyzing the results. Thus, to explore the effect of factoids on getting the question right, we built logistic regression models using SPSS. Logistic regression predicts a binary outcome variable using several categorical factors, and is a statistically valid technique for analyzing experiments with multiple responses per student – more sensitive than analysis of variance over the mean of students' answers, and more statistically appropriate than considering each separate answer as an independent trial.[1] Here, the outcome variable was whether the student got the answer right or not. The following factors were

[1] See glossary for details.

included in the model:

- whether student received a factoid on the target word,
- who the student was, so as to prevent bias towards students with more trials, and to properly group a student's trials together;
- a term for how difficult the questions were overall – that is, background difficulty – and
- what the effect of help was on getting the question right.

If the coefficient for the effect of help on getting the question right was (significantly) greater than zero, then factoids (significantly) boosted student performance. We accompany the description of results below with figures on average percent correct, calculated on a per-student basis to avoid bias towards students who encountered more target words

4.4.1 Overall, factoids did not help…

Did factoids significantly boost performance? The per-student average percentage correct for the control trials was 37.2% ± s.d. 16.9%; for experimental trials, 38.5% ± s.d. 18.3%. (Per-student percentages have high standard deviations because they are averages of individual rates, which vary by student.) The coefficient for the effect of help on getting the question right was 0.07 ± 0.07, for all 3359 trials. Thus, factoids did not significantly boost performance overall – due perhaps to a number of problems with automated assistance that we next sought to filter out.

4.4.2 Exploration revealed possible effect for rare single-sense words, tested one or two days later

We decided to examine conditions under which the factoids might have been effective. Therefore, we looked at subsets of the data to see how factoids helped – or not – under various conditions. The exploratory nature of the following analysis means that its results should be considered suggestive, not conclusive. What conditions might affect the effect of factoids?

Some words in the target set had more than one meaning. Students might well be confused – or at least not helped – by factoids that explained a different sense of the target word than was used in the original text. Perhaps factoids were effective only for single-sense words. Did factoids help for single-sense words only? Not significantly, but the trend was still positive (Table 4.3).

Some of the words in the target set were easy – *apple*, for example. Presumably, if a student already knew a word, a factoid would not help. Did factoids help for single-sense hard words? Maybe. We manually classified each target word as hard or not hard. So as to avoid biasing the classification due to knowledge of the outcome of the trials, we classified the words without looking at the outcomes on individual trials or words. We also identified the words that were rare – words that occurred fewer than 20 times in the million-word Brown corpus (Kucera and Francis 1967). Results were again not significant, but suggestive of a positive impact of factoids, as follows (Table 4.3).

Perhaps students learned or remembered enough of the help to do better a few days later, but not over an extended period of time such as a weekend. Did the factoids help for single-sense rare words tested one or two days later? Yes (Table 4.3). If the effect only persists for a few days, how could we improve students' retention of the meanings they learned? Future work might aim at reinforcing this retention with a second exposure to the target word.

Number of trials	How were trials selected?	Per-student average number right	Coefficient in logistic regression model
720 trials	Single-sense words	34.9% ± 23.0% for control vs. 38.4% ± 26.5% for experimental	0.23 ± 0.17
191 trials	Single sense words coded as hard by a certified elementary teacher	26.3% ± 30.0% for control vs. 29.1% ± 36.8% for experimental	0.13 ± 0.41
348 trials	Single sense words coded as hard by the experimenter	33.0% ± 29.0% for control vs. 40.7% ± 36.3% for experimental	0.35 ± 0.27
317 trials	Single sense rare words	35.4% ± 30.5% for control vs. 42.4% ± 37.3% for experimental	0.16 ± 0.29
189 trials	Single-sense rare words tested one or two days later	25.8% ±29.4% for control vs. 44.1% ± 37.7% for experimental	1.04 ± 0.42 Significant at 95%, exploratory and thus not correcting for multiple comparisons

Table 4.3. Single-sense difficult words in the factoids experiment.

Word	Word frequency in Brown corpus	Example of a factoid	Example of a multiple-choice question – correct answer in **bold**, student’s response underlined.
aluminum	18	aluminum can be a kind of metal	**Al**;wood coal;soot;black lead
astronaut	2	astronaut can be a kind of traveler	married person; **traveler**; computer; nerd
bliss	4	Maybe bliss is like walking on air here	seeing red; scare; **walking on air**; melancholy
bobbin	-	Maybe bobbin is like reel here	**reel**; wheel; power train; gun
coward	8	coward can be a kind of mortal	Old Nick; young; **someone**; escape
crouching	-	crouch can be a kind of sit down	**crouched**; lace; wring; mat
daisies	-	daisy can be a kind of flower	prairie star; painted cup; snow plant; **flower**
eggshell	1	eggshell can be a kind of natural covering	Little Dog; mouth; meat; **cover**
glittering	-	-	stay together; **shine**; carry; lurch
headdress	-	headdress can be a kind of apparel	plastic wrap; **wearing apparel**; dust cover; arm
hello	10	Maybe Hello is like hi here	good day; good night; morning; **hi**
infirmities	-	infirmity can be a kind of bad condition	sore; **bad condition**; wound; twist
liar	3	liar can be a kind of cheat	runner; **cheat**; beast; wolf
outskirts	-	outskirt can be a kind of city district	hub; nation; **city district**; roads
pasta	-	pasta can be a kind of food product	**food product**; chow; bird food; food cache
pebbles	-	pebble can be a kind of stone	clay; **rock**; sheath; Crow
plat[2]	-	plat can be a kind of map	**map**; chalk; check; rule
plumage	-	plumage can be a kind of animal material	mineral pitch; **body covering**; dye; winter’s bark
pollen	11	Pollen can be a kind of powder	**powder**; diamond dust; water glass; milk glass
princess	10	Princess can be a kind of blue blood	coach; chair; **blue blood**; mayor
Rwanda	-	Rwanda can be a kind of African country	**African country**; England; United Kingdom; United States
salad	9	salad can be a kind of dish	bite; **dish**; breakfast; choice morsel
tennis	15	tennis can be a kind of court game	field game; night game; **court game**; day game
twinkling	2	-	ping; **second**; ring; bang
vales	-	Maybe vale is like valley here	Blue Ridge Mountains; hill; **valley**; bank
wading	-	wade can be a kind of walk	work; pace; bounce; **walk**
wayside	2	wayside can be a kind of edge	arm band; **margin**; strap; ring

Table 4.4. Single-sense rare words tested one or two days later.

[2] *plat* appeared in an (apparently) student-written story which included the sentence “slapt flash slise plair clio

As a sanity check, we looked at the 27 words in these trials (Table 4.4). Word frequency in Table 4.4 is out of the million-word Brown corpus (Kucera & Francis 1967). *glittering* and *twinkling* turned out not to have been supplemented with factoids during the study – due to the random assignment of trials to conditions. Most of the words in Table 4.4 seem plausible as words that some elementary school students might not know, and for which explanations might be helpful. Selecting trials where the test occurred only one or two days after the training meant including fewer trials from students who were frequently absent, introducing a self-selection bias. Therefore, we next explored the factoid results using attributes that did not reflect self-selection, but rather other properties of the students such as grade.

4.4.3 Further characterization of factoid results

In order to more fully characterize the factoid results – to find where the factoids might have helped – we looked at a number of possible subdivisions of the data with respect to their effects both on the percentage of correct answers, and on the coefficient for effect of factoid on answer in the regression model. Table 4.5 shows percentage correct – calculated as the average of the per-student mean – and the effect of factoid on answer for several subdivisions of the data.

ciay glass plat”.

Which students?	Which words?	Trials	Percentage correct	Outcome: Coefficient ± 1 s.d.
All students	All words	3359	37.2% ± 16.9% control 38.5% ± 18.3% expt.	No significant effect: 0.07 ± 0.07
33 students in Grade 2	All words	1391	35.4% ± 11.7% control 33.1% ± 11.6% expt.	No significant effect: -0.03 ± 0.12
36 students in Grade 3	All words	1968	33.1% ± 11.0% control 42.0% ± 19.0% expt.	Trend favoring factoid: 0.15 ± 0.10
All students	Single-sense	769	36.8% ± 26.6% control 39.2% ± 29.2% expt.	Slight trend favoring factoid: 0.21 ± 0.17
All students	Multiple-sense	2605	37.4% ± 17.2% control 37.3% ± 20.2% expt.	No significant effect: 0.07 ± 0.08
All students	Rare words	1927	35.6% ± 19.5% control 38.3% ± 21.1% expt.	No significant effect: 0.13 ± 0.10
All students	Non-rare words	1427	40.0% ± 18.6% control 37.8% ± 22.3% expt.	No significant effect: -0.06 ± 0.11
Grade 3	Rare words	465	36.2% ± 22.9% control 42.0% ± 28.4% expt.	Effect of factoid: .37 ± .21, p < .10
29 students below median on weighted score of WRMT word comprehension pretest	All words	1319	33.1% ± 11.1% control 32.9% ± 9.4% expt.	No significant effect: 0.07 ± 0.12
31 students at or above median on weighted score of WRMT word comprehension pretest	All words	1852	38.3% ± 9.7% control 42.3% ± 16.6% expt.	No significant effect: .11 ± .10

Table 4.5. Further characterization of factoid results.

4.4.4 Word recency effect

The comparison word (*traveler* in “*astronaut* can be a kind of *traveler*”) and the expected correct answer were drawn from partially overlapping sets of words. Because of the overlap between sets, 993 out of the 1709 experimental trials in this experiment used the same word for the comparison word and the expected answer, and the other 716 used a different word. The effects found when analyzing all of the trials could be due solely to a recency effect from having seen the comparison word on a previous day. Later experiments on augmenting text with definitions were designed to avoid such recency effects.

4.5 Lessons learned from factoid study

In conclusion: The factoid study suggests that augmenting text with factoids can help students learn words, at least for third graders seeing rare words (p < .10), and for single-sense rare words tested 1-2 days later (p < .05).

There were various problems with factoid assistance that may have diluted the effectiveness of factoids. We have already discussed target word frequency, multiple senses, and socially unacceptable factoids or test items. In addition, other problems remained. For the vocabulary assistance, some comparison words may have been harder to understand than the target words.

There were also various problems with the automated assessment that may have obscured the effectiveness of factoids.

1. For example, some of the incorrect answers (distractors) were themselves rare – such as *butterfly pea* – making the question difficult to understand.
2. Or, questions may have relied on uncommon knowledge, such as *banana* being (botanically) an *herb*.

These examples illustrate the absence of some constraints that common sense would enforce, but that a computer program must explicitly allow for.

In fact, at the end of Fall 1999, we turned off the vocabulary questions primarily because they were getting slower and slower as database queries labored over data collected during the entire year to date, but also due to the problems we just discussed. We did however leave the factoids on, to avoid excessive changes to what children did on the Reading Tutor. While turning vocabulary questions off precluded carrying out fine-grained analysis of factoids in Spring 2000, in Chapter 5 we present a more summative analysis: a year-long evaluation of the Reading Tutor with Take Turns and factoids, which compared the Reading Tutor to classroom instruction, and also to one-on-one human tutoring.

In summary, then, factoids helped – sometimes – but generating good assistance automatically requires common sense that code lacks. Thus, at least for the near term we recommend using vocabulary assistance as follows: either constructed by machine and then hand filtered, or directly constructed by hand. (We followed our own advice, as the designs of experiments reported in Chapter 6 bear out.)

5 How well did the 1999-2000 Reading Tutor help children learn vocabulary?

So far we have described two enhancements to computer-assisted oral reading. Chapter 3 described how we changed the Reading Tutor's story choice policy to have the computer and the student take turns picking stories. We showed that Take Turns resulted in students reading more new material than they presumably would have on their own. Chapter 4 discussed how we enriched text with vocabulary assistance in the form of automatically generated factoids like "astronaut can be a kind of traveler. Is it here?" We showed that factoids helped, at least for third graders seeing rare words, and also for single-sense rare words tested one to two days later.

So, the changes we made improved the baseline 1997-98 Reading Tutor. But how did the new and improved Reading Tutor with Take Turns and factoids compare to other methods of helping children learn to read? Specifically, how did the 1999-2000 Reading Tutor compare to other reading instruction, on measures of vocabulary learning? In this chapter we present relevant parts of a larger 1999-2000 study that we helped design and analyze, but that was carried out primarily by other Project LISTEN team members. This larger study was not intended solely to evaluate vocabulary assistance, but did include vocabulary gains in comparing the modified Reading Tutor against other treatments.

Project LISTEN conducted a year-long study in 1999-2000, comparing three treatments: the Reading Tutor, human tutoring in the form of assisted reading and writing, and a control condition with equal-time classroom instruction (not necessarily in reading). The primary purpose of the year-long study was to "prove and improve" the Reading Tutor: to compare the

Reading Tutor to conventional methods of reading instruction, and to identify areas for improvement. This dissertation focuses on vocabulary learning; therefore we report here only the parts of the story relevant to vocabulary learning.

In the remainder of this chapter, we describe the students in the study, specify the treatment conditions, report how learning outcomes were measured, present results, and draw implications. Appendix A gives details on the study according to the National Reading Panel's schema (NRP 2000); here we provide a brief summary. See Aist et al. (AI-ED 2001) and Mostow et al. (AI-ED 2001) for details.

5.1 Assignment of students to treatments

A total of 144 students in grades 2-3 at Centennial Elementary School, an urban elementary school near Pittsburgh, Pennsylvania, participated in the 1999-2000 study. Six of twelve (12) classrooms were assigned to receive a Reading Tutor, based on the school principal's estimate of teachers' willingness to cooperate. This possible confound is partially ameliorated as follows: according to the principal, all classroom teachers in the study were comparably experienced veteran teachers. (Randomized assignment of rooms to conditions would have produced uninformative results if it had led simply to poor treatment fidelity.) The classrooms used a basal reading curriculum. Class size was approximately 24 students. In each Reading Tutor classroom, the teacher identified the 12 lowest-reading students. Of those 12, we randomly picked one from the lower 6 and one from the higher 6 as within-room controls, and assigned the other 10 to use the Reading Tutor for 20 minutes every day, which we thought would be difficult to do for more than 10 students per room. With our blessing, teachers tried to get 3 of the within-room control students on the Reading Tutor as well, but indeed were unable to get them on the Reading Tutor for the same amount of time. We therefore excluded these 3 students from the analysis. In each non-Reading Tutor classroom, the lowest-reading 12 students were assigned randomly (with pairing) to receive either baseline classroom instruction or one-on-one human tutoring.

Assigning the students to treatments in this way means that students in different conditions may have received different amounts of instruction on various reading skills. That is, the 20 minutes of treatments in each condition may have been allocated differently among different reading (and writing) skills, and in the classroom control condition, between reading, writing, and other instruction.

5.2 Treatments

We describe each treatment condition in turn.

Classroom control (labeled “control” in figures). Students in the baseline condition received normal instruction.

Human tutoring (labeled “human” in figures). Students in the human tutoring condition received nominally 20 minutes of tutoring per day, at a desk in the hallway just outside the classroom. One human tutor was assigned to each control classroom, except for occasional substitutions due to human tutors’ other responsibilities.

Studies of one-on-one human tutoring for elementary reading have employed tutors with varying degrees of training, from volunteers (Juel 1996) to paraprofessional teachers’ aides to certified teachers to certified teachers with specialized training in a particular reading program. Wasik and Slavin (1996) found that using certified teachers rather than paraprofessionals was associated with positive results for one-on-one reading tutoring. In this study, the human tutors were already employed by the school as tutors. The human tutors were certified elementary teachers with at least a bachelor’s degree in elementary education and 0-2 years teaching experience (in many cases involving preschool children), but with no specialized training in reading tutoring. Thus the tutors could be expected to do better than classroom instruction, and provided a realistic comparison condition that could be replicated on a larger scale – unlike hiring the world's best tutor.

During tutoring sessions, human tutors helped students with reading and writing. To control for the important factor of materials, the human tutors used paper copies of the Reading Tutor stories, and were asked not to bring in outside material. Human tutors were however free to engage in other activities (such as individual word review) so long as they kept written records of the activities. Human tutors also kept track of which stories their students read in a "Human Tutor Log", and had the students write (when they did so) in a journal. Project LISTEN collected the logs and the journals at the end of the year, and for convenient capture and timely access used a digital camera to record them on-site during the year.

Reading Tutor (labeled "computer" in figures). There was one Reading Tutor computer per Reading Tutor classroom in the 1999-2000 study. Students in the Reading Tutor condition were scheduled for 20 minutes of Reading Tutor use, working individually.

Tutoring in the human tutor and Reading Tutor conditions took only a small portion of the school day. Thus most of the instruction that students received was the same for students in the same classroom, regardless of treatment condition. Teachers rotated the scheduling of Reading Tutor time and human tutor time to keep students from consistently missing the same subject. Mostow et al. reported regarding time on task, "We assigned each student to the same treatment 20 minutes daily for the year. Each day included 60-70 minutes of reading instruction plus varying time on related activities. Thus all students in a given classroom received mostly the same instruction, with tutored students averaging 0 to 15 minutes more time per day on reading and writing. Teachers rotated scheduling to vary which subjects students missed while being tutored" (Mostow et al. AI-ED poster 2001).

All three treatment conditions included a range of activities, including some directed at vocabulary development. Thus in this chapter we are comparing three comprehensive treatments on a single aspect of learning to read, not three treatments aimed specifically at encouraging vocabulary development. Vocabulary development alone might be accomplished more efficiently by a more narrowly targeted treatment – presumably by sacrificing the rich experience of encountering words in context, replacing it with some task such as pairing words with synonyms or definitions.

Out of the 144 students who began the study, 131 completed the study. Table 5.1 shows the division of students into conditions.

Room * Treatment * Room type (Control vs. Reading Tutor * Grade Crosstabulation						
				Treatment		
Grade	Room type			control	human	computer
2	Control	Room	205	6	6	
			208	6	5	
			209	3	6	
	Reading Tutor	Room	201	2		9
			211			10
			212	2		10
3	Control	Room	208		1	
			305	5	5	
			309	5	6	
			310	6	5	
	Reading Tutor	Room	301	2		8
			303	2		10
			304			11
Totals by grade				Grade 2: 19 Grade 3: 20	Grade 2: 17 Grade 3: 17	Grade 2: 29 Grade 3: 29

Table 5.1. Assignments of students to treatments, by grade and classroom, showing the 131 students who completed the study. (One student – in room 208 here – switched grades from Grade 2 to Grade 3.)

5.3 Outcome measures

To gather results comparable to other published studies on reading instruction, Project LISTEN used the Woodcock Reading Mastery Test (WRMT) (American Guidance Service, n.d.), an individually administered reading test normed by month within grade to have a mean of 100 and a standard deviation of 15. The WRMT consists of several subtests, each of which tests a specific area of reading skill. In this study, trained testers pre- and post-tested students using the following subtests of the Woodcock Reading Mastery Test: Word Attack (decoding skills), Word Identification (reading single words out loud), Word Comprehension (single word understanding), and Passage Comprehension (understanding 1-2 sentence passages). The testers also measured students' oral reading fluency: their unassisted oral reading rate on prespecified passages at grade level and at a student-appropriate level.

5.4 Results on Word Comprehension

Because this dissertation focuses on vocabulary learning, we discuss here only the results for the Word Comprehension subtest. The analyses in this chapter were principally conducted by the present author and Project LISTEN team member Brian Tobin, with additional advice from others. We addressed several questions, as follows. Did the children working with the Reading Tutor:

1. gain from pre- to post-test?
2. gain more than a national cohort?
3. gain more than their peers who received classroom instruction?
4. gain more than their peers who received one-on-one human tutoring?

We look at each question in turn.

5.4.1 Did Reading Tutor students gain from pre- to post-test?

Yes: the difference between post-test and pre-test "raw" weighted score (prior to norming) on Word Comprehension was 15.72 ± standard error 1.12, with 95% confidence interval (13.49, 17.96). Figure 5.2 shows boxplots of students' raw gains on Word Comprehension, by treatment groups; Figure 5.3, by treatment and grade. This gain is not all that interesting because it might simply reflect children's general growth over the year; to filter out such general growth, we next compared to the national norm.

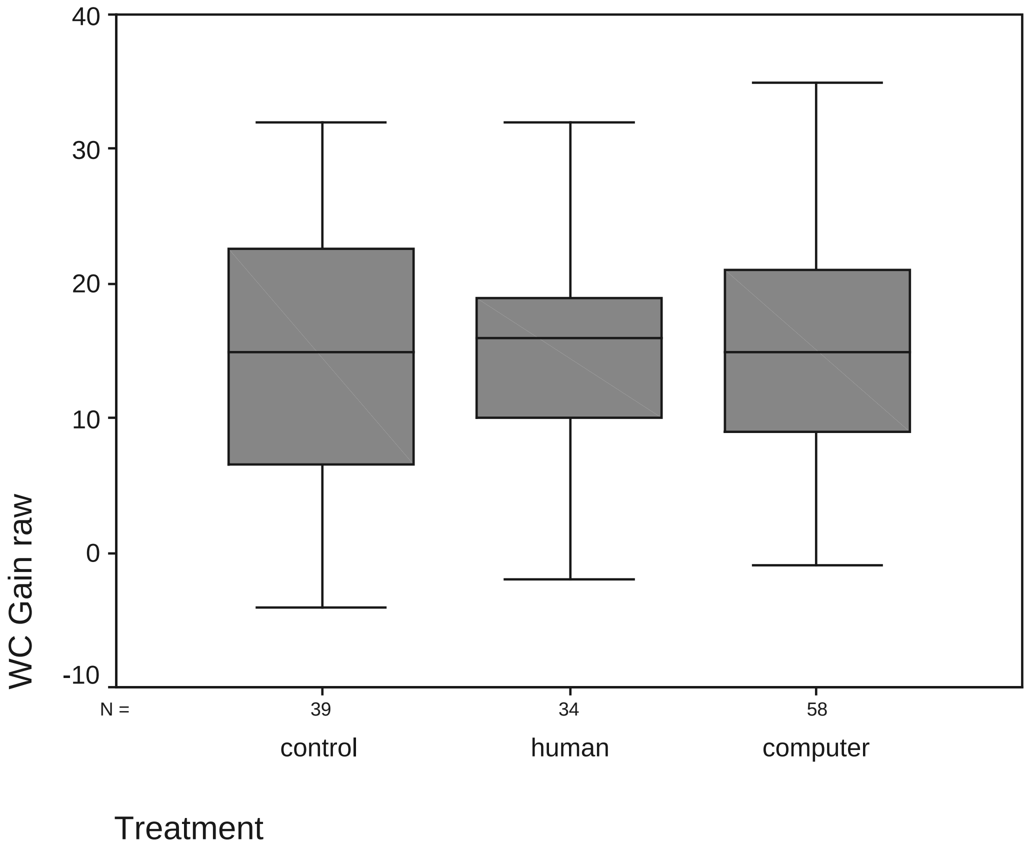


Figure 5.2. Word Comprehension raw score gains, by treatment. 1999-2000.

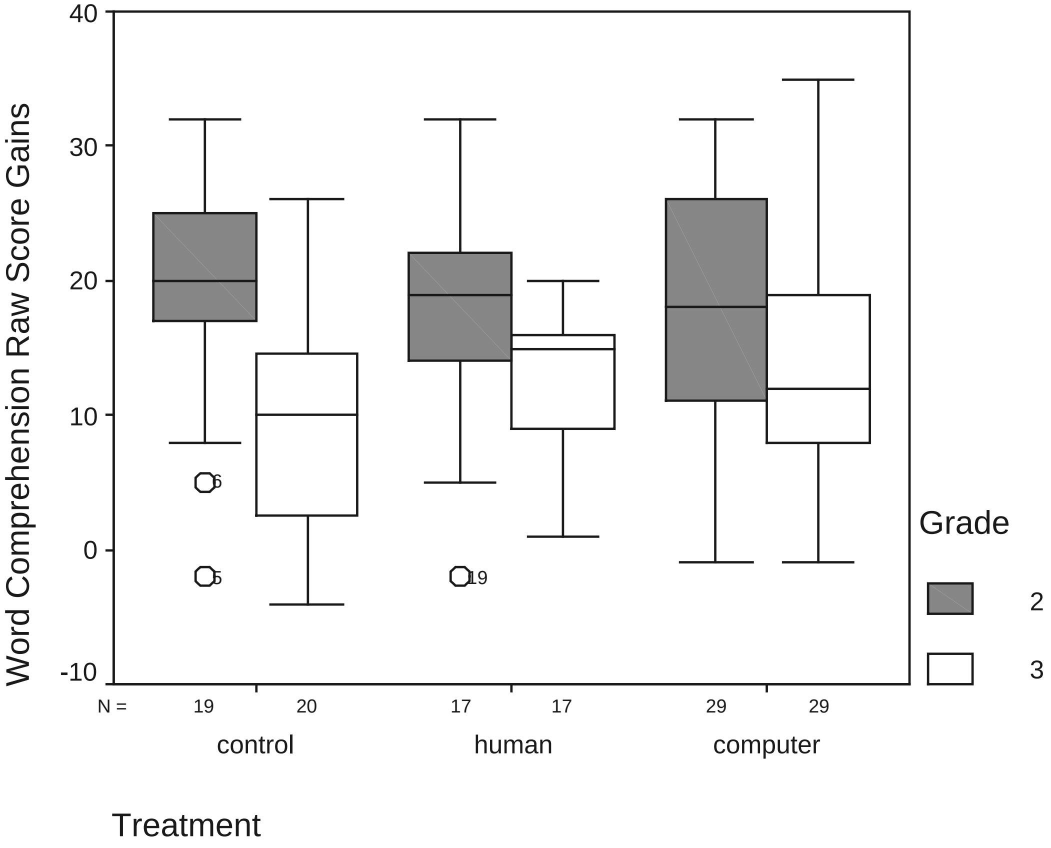


Figure 5.3. Raw score Word Comprehension gains by grade and treatment.

5.4.2 Did Reading Tutor students gain more than a national cohort?

To answer this question, we looked at the normed gains. (A gain of zero on a normed score means that a student stayed at the same level from pre-test to post-test relative to the norming sample – not that he or she learned nothing, but that he or she learned enough to stay at the same level with respect to the norms.) Students who read with the Reading Tutor averaged gains of 4.38 ± standard error 0.90, with 95% confidence interval (2.58, 6.18). Therefore students who read with the Reading Tutor learned enough to move forward with respect to the normed scores. Figure 5.4 shows boxplots of normed score gains for treatments; Figure 5.5, for grade and treatment. Figures 5.6 and 5.7 show normed pretest, post-test, and gains by grade.

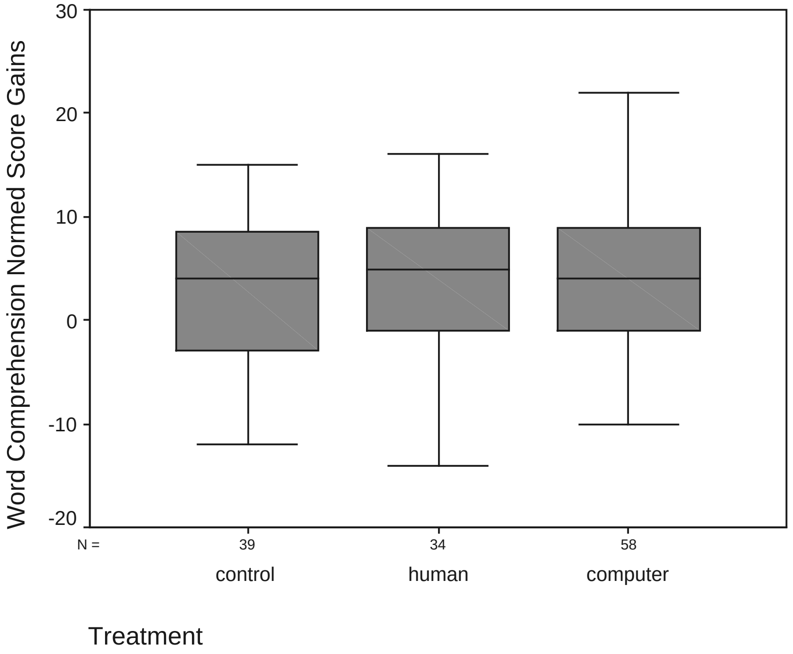


Figure 5.4. Word Comprehension normed score gains, by treatment. 1999-2000.

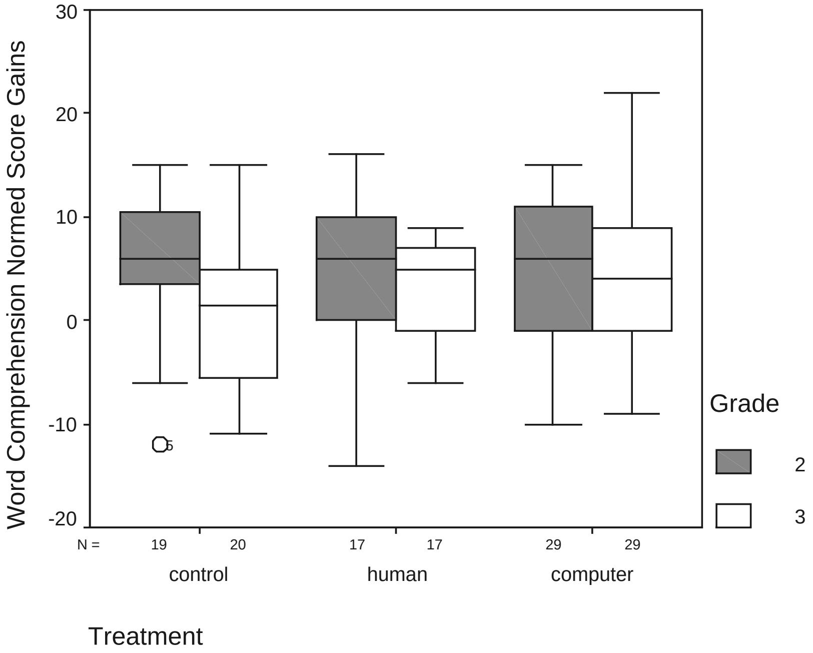


Figure 5.5. Word Comprehension normed score gains, by grade and treatment.

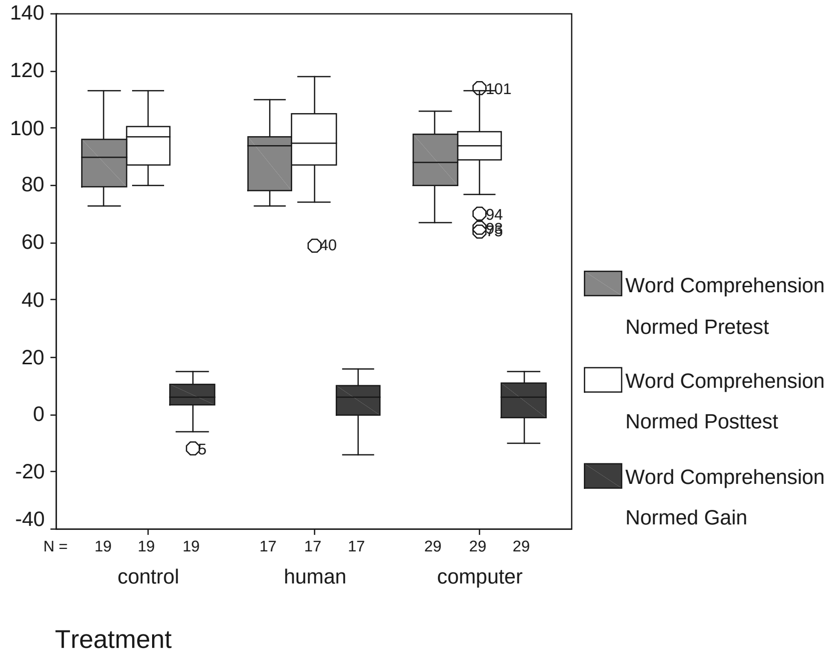


Figure 5.6. Word Comprehension normed pretest, post-test, and gains, Grade 2.

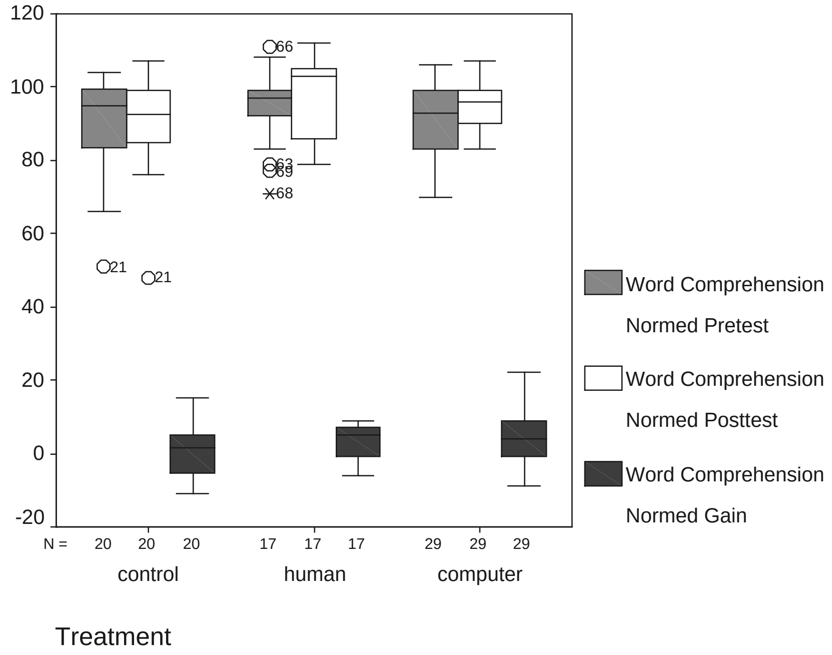


Figure 5.7. Word Comprehension normed pretest, post-test, and gains, Grade 3.

5.4.3 Did Reading Tutor students gain more than their peers who received classroom instruction?

For the overall results, we used analysis of variance of Word Comprehension gains by treatment and grade, with an interaction term for grade and treatment. Table 5.2 shows results by room-treatment pairs; since there were six separate human tutors, Table 5.2 presents results for each human tutor individually, with the Reading Tutor separated out by room for comparison purposes. Table 5.3 presents an overall summary. To control for regression to the mean, and other differential effects of pretest on gain, we included (normed) Word Comprehension pretest. We included (normed) Word Identification pretest to maximize the fit of the model to the data, arrived at by searching through the set of combinations of possible covariates (Word Attack, Word Identification, Word Comprehension, Passage Comprehension, and fluency) and minimizing the error remaining between the model and the actual data. Analysis of variance using raw scores and considering both grades together revealed a significant interaction between treatment and grade (F=2.47, p=.088), so we considered grade 2 and grade 3 separately.

Considering grades 2 and 3 separately, there was no significant effect for treatment in grade 2 (F=0.32, p = .731). However, all of the treatment groups appeared to improve – even in comparison to the national norm. Perhaps the several hours per day of classroom instruction at this school (at least in grade 2) were effective enough at building children’s vocabulary that the classroom gains masked whatever benefit the human tutors – and the Reading Tutor – provided. In grade 3, there was a significant main effect for treatment (F=4.27, p =.018; significant at 95% even if applying Bonferroni correction because there were only two grades in the study). The students on the Reading Tutor in grade 3 did better than their peers receiving classroom treatment, with an estimated advantage on Word Comprehension normed score gains of 3.90

points ± standard error 1.54; for grade 3, effect size[3] = 0.56, and p = .042, with Bonferroni correction for multiple pairwise comparisons. So, third graders using the Reading Tutor did better than their peers receiving classroom instruction.[4] (Third graders receiving human tutoring likewise did better than their peers receiving classroom treatment, at 4.56 ± standard error 1.78, effect size = 0.72, p = .039 with Bonferroni correction). Incidentally, showing gains vs. a control group precludes regression to the mean as an explanation of the effect – because the control group would presumably show such regression as well.

We continued to use this analysis to answer our fourth question.

5.4.4 Did Reading Tutor students gain more than their peers who received one-on-one human tutoring?

Again, there was no main effect for treatment in grade 2. In grade 3, there was no significant difference between the human tutored students and the students who read with the Reading Tutor (0.66 points more on normed Word Comprehension gains in favor of the human tutored students, with standard error ± 1.65). How did the Reading Tutor compare to human tutors? Table 5.3 shows how the Reading Tutor compared to individual human tutors and to classroom instruction on gains for Word Comprehension disaggregated by grade.

[3] Effect size is the adjusted gains difference divided by the average standard deviation of the compared subtests.

[4] Incidentally, gains vs. a control group does not permit regression to the mean as an explanation of the effect.

	Grade 2	Grade 3
Individual results normed by grade	9.67 ± 2.42 HT Room 205 n=6	9.50 ± 6.36 CT Room 303 n=2
	7.83 ± 6.31 CT Room 208 n=6	7.60 ± 5.13 CT Room 305 n=5
	7.00 ± 2.65 CT Room 209 n=3	**6.00 ± 5.18 RT Room 301 n=8**
	6.17 ± 6.71 HT Room 209 n=6	4.80 ± 5.17 HT Room 305 n=5
	5.70 ± 7.45 RT Room 212 n=10	**4.36 ± 7.23 RT Room 304 n=11**
	5.00 ± 0.00 CT Room 212 n=2	3.83 ± 4.31 HT Room 309 n=6
	4.33 ± 10.82 CT Room 205 n=6	**2.90 ± 6.38 RT Room 303 n=10**
	3.90 ± 7.92 RT Room 211 n=10	2.20 ± 6.76 HT Room 310 n=5
	3.67 ± 7.68 RT Room 201 n=9	1.00 ± 5.66 CT Room 301 n=2
	1.50 ± 6.36 CT Room 201 n=2	-4.00 ± 4.98 CT Room 310 n=6
	-3.50 ± 6.66 HT Room 208 n=6	-4.40 ± 5.08 CT Room 309 n=5
Individual results normed by grade, from a model including Word Comprehension pretest and Word Identification pretest as covariates	9.81 ± 2.48 HT Room 205	7.80 ± 4.33 CT Room 303
	6.91 ± 2.49 CT Room 208	6.29 ± 2.74 HT Room 305
	6.56 ± 3.54 CT Room 209	5.92 ± 2.75 CT Room 305
	6.37 ± 4.46 CT Room 212	**5.35 ± 1.85 RT Room 304**
	6.25 ± 2.48 HT Room 209	**5.29 ± 2.16 RT Room 301**
	4.77 ± 1.93 RT Room 212	4.21 ± 2.76 HT Room 310
	4.37 ± 2.48 CT Room 205	4.04 ± 2.48 HT Room 309
	3.54 ± 2.04 RT Room 201	**2.46 ± 1.93 RT Room 303**
	2.76 ± 1.97 RT Room 211	1.57 ± 4.30 CT Room 301
	1.61 ± 4.33 CT Room 201	-3.08 ± 2.50 CT Room 310
	-2.22 ± 2.52 HT Room 208	-3.96 ± 2.73 CT Room 309

Table 5.2. Word Comprehension normed score gains by grade, for classroom control (CT), human tutors (HT), and the Reading Tutor (RT). Results prior to covarying out pretest show mean plus or minus standard deviation; results after covarying out Word Comprehension and Word Identification show mean plus or minus standard error.

Table 5.2 shows results by grade. For second graders, all three treatment groups achieved approximately equivalent gains. For third graders, the Reading Tutor-ed students gained more than the classroom instruction group on Word Comprehension, and the human-tutored students likewise gained more than the classroom instruction group. How did the Reading Tutor compare to human tutoring? Overall, the Reading Tutor and the human tutors were comparable. Small sample sizes for the human tutors preclude a statistically precise ranking of all the human tutors and the Reading Tutor. Nonetheless, the Word Comprehension gains (both as-is, or adjusted by covariates) by the students in the study placed the Reading Tutor groups roughly interspersed with the human tutored groups.

Were there significant differences among the human tutors for Word Comprehension gains? Yes (Table 5.3). For the second graders, the students tutored by the human tutor M.B. achieved significantly greater gains on Word Comprehension (9.81 ± standard error 2.48) than the students tutored by the human tutor M.E. (-2.22 ± standard error 2.52). To further improve the Reading Tutor on vocabulary learning, we may in the future look at M.B.'s tutoring strategies and compare them to M.E.'s strategies.

	Overall	Grade 2	Grade 3
Compare classroom-tutored students, human-tutored students, and students who used the Reading Tutor	Marginally significant effect of grade*treatment interaction: F = 2.47, p = 0.088.	No significant main effect of differing treatment conditions	Significant main effect for treatment: F = 4.268, p = 0.018. Human tutor > Classroom: HT – CT = 4.56 ± 1.78 p =0.039 with Bonferroni correction for multiple comparisons Reading Tutor > Classroom: RT – CT = 3.90 ± 1.54 p = 0.042 with Bonferroni correction Reading Tutor ≈ Human tutor: RT – HT = -.663 ± 1.646, p = 1.0
Compare individual human tutors and the Reading Tutor	Marginally significant main effect of different tutors: F=1.8, p = 0.098	Marginally significant main effect of different tutors: F=2.64, p = 0.063	Different tutors not significantly different
Tutors ordered from highest to lowest grade-normed word comprehension gains		Human tutor MB: 9.7 ± 2.4, 6 students Human tutor AC: 6.2 ± 6.7, 6 students Reading Tutor: 4.4 ± 7.5, 29 students Human tutor ME: -4.0 ± 7.3, 5 students	Human tutor LN: 4.8 ± 5.2, 5 students Reading Tutor: 4.3 ± 6.3, 29 students Human tutor MM: 3.1 ± 4.3, 7 students Human tutor NJ: 2.2 ± 6.8, 5 students

Table 5.3. Comparing the Reading Tutor to human tutoring, classroom instruction.

5.5 Relationship between students' word comprehension gains and distinct words seen

So students gained in word comprehension with the Reading Tutor – and third graders gained more than their peers in a classroom-instruction control condition. Remember that our informal model of vocabulary learning predicts that the number of words seen is positively related to the number of words learned (Equation 1.1, Chapter 1; Equation 2.1, Chapter 2). How do students' gains compare to our informal model of vocabulary learning?

To determine whether students' encounters with words significantly affected their word comprehension gains, we calculated the partial correlation between a student's grade-normed

Word Comprehension gains and the number of distinct words that particular student saw in the Reading Tutor (see Figure 3.12). Word Comprehension pretest scores were significantly correlated with gains (-.4034, p=.002) and – almost – with words seen (.22, p=.102), so when calculating the partial correlation we correlated out the grade-normed Word Comprehension pretest. After controlling for grade-normed Word Comprehension pretest, the partial correlation between grade-normed Word Comprehension gains and distinct words seen in the Reading Tutor was .18, p = .178. This correlation is interesting but not very strong. Furthermore this correlation does not entail causality; there could be a common cause for both, such as differences in students' motivation to select appropriately challenging material.

We defer to future analysis the problem of assigning credit for these gains to story choice (Chapter 3) or the factoids (Chapter 4). Nonetheless, this result provides the first direct evidence suggesting not only that our model of vocabulary learning reflects students' learning outcomes, but also that seeing new words in the Reading Tutor was directly related to students' vocabulary development – above and beyond what could be explained solely by their previous vocabulary.

5.6 Lessons learned

Results from this year-long study are as follows: a computer tutor that did better than a classroom control for third graders' vocabulary learning. The 1999-2000 Reading Tutor with Take Turns and factoids even performed competitively with one-on-one human tutoring. But how can we push forward and make reading with the Reading Tutor even more effective for vocabulary acquisition? In Chapter 6 we describe experiments to further clarify which students benefit when from vocabulary assistance.

6 Follow-on experiments in vocabulary assistance

In this chapter we discuss two follow-on experiments on a promising direction for vocabulary help: in-context explanations. First, we checked to make sure that students in this chapter's intended subject pool (low-reading elementary students) could understand and make immediate use of the information in explanations. Second, we conducted an experiment to tease apart the effects of seeing a word in running text versus seeing a word in an explanation. We discuss each experiment in turn.

6.1 Can (low-reading elementary) students make use of explanations?: The comets and meteors experiment

How can we construct vocabulary assistance that is even more effective than comparisons to other words, even if it takes a bit longer? Pictures may help for some words, but can't easily illustrate all words – for example, *mendacious* (untruthful) is easy to explain, but hard to draw. What about having students read definitions from a children's dictionary? Conventional dictionary definitions have been shown to be less effective at teaching word meanings than context-specific definitions (McKeown 1993). Therefore, we explored inserting short, context-specific explanations into text.

If students are to learn from explanations, they must first understand them. The fact that factoids' effectiveness was restricted to only certain subsets of students and words, as described in Chapter 4, made us concerned that experiments with explanations might not find any significant effects on learning. Thus we first conducted an experiment to check whether in fact our intended population – low-reading elementary students, 2nd through 5th grades – could understand short explanations of words. We conducted an experiment on paper to make sure that students could understand explanations. We compared short explanations to nonsemantic assistance ("COMET starts with C"), to ensure that any advantage was due to understanding the content of the explanation, not simply seeing another exposure. In each case, the extra sentence was presented just prior to the sentence from the original text containing the target word. The test was a five-item matching task. We now describe in more detail the texts, the test, the experiment design, the results, and some implications.

6.1.1 Texts

We adapted two texts from the StarChild website, a science website for children (http://starchild.gsfc.nasa.gov/docs/StarChild/StarChild.html). Middle-school teachers wrote both texts. The texts were nonfiction; one text was on comets and the other text was on meteors. For example, the opening sentence of the comets story was "Scientists believe that comets are made up of material left over from when the Sun and the planets were formed." To achieve approximate equality in length, we edited the two texts by deleting sentences from one passage and adding text to the other from the version of the text aimed at older students. Our edits also aimed at preserving the gist and flow of the passage. The complete text of each story, original and edited, is given in Appendix D.

The original Web-formatted texts contained hyperlinks to an online glossary written by the same teachers who wrote the original text. For example, the definition for *comet* was "COMET:

A big ball of dirty ice and snow in outer space." For our paper version, we interspersed either these explanations or nonsemantic assistance into the original text. For example, the nonsemantic assistance for *comet* was "COMET starts with C." The definitions and the nonsemantic assistance are both given in full in Appendix D.

The original Web pages also contained pictures: a floating girl reading; a boy holding a toy spaceship; comets orbiting the sun; and meteors falling through the night sky. While the Reading Tutor can (and sometimes does) display pictures that illustrate the story, in our experiment we were testing text only (without pictures), so we omitted the pictures, and presented the texts on paper in a large, child-friendly font.

Table 6.1 presents a summary of the texts. (Grade levels shown are only approximate, as readability formulas simply estimate grade level of text.)

	Adapted text	Text plus nonsemantic help	Text plus definitions
Comets	168 words Grade level 5.7	189 words Grade level 4.9	241 words Grade level 6.3
Meteors	173 words Grade level 6.1	194 words Grade level 5.4	261 words Grade level 6.1

Table 6.1. Summary of texts used in comets and meteors experiment. Grade levels calculated using Flesch-Kincaid Grade Level using Microsoft Word (Office 97, Windows NT).

6.1.2 Test

We measured understanding of the target words in the story as follows. The test consisted of one 5-item matching task for each topic. The tests were administered the same day – in fact, the tests were stapled to the two stories. For example, the match for the word *comet* was "A ball of ice in space with a tail." The matching items did contain some lexical overlap with the in-text definitions (cf. Chapter 4) – writing a definition for *comet* without mentioning *ball*, *ice*, *space*, or *tail* would most likely result in a forced, unnatural-sounding definition. The matching items and the definitions were however different at the phrasal level. Because students could flip back and forth between the stories and the tests, the paper experiment tested a combination of short-term memory, reading comprehension, and information access. The matching tasks for the comets and meteors stories are given in full in Appendix D.

What was the chance-level performance on the matching test? Because the test was a matching task, performance on items was not independent: answering one item affected students' choices on the others. That is, once a student had decided how to match one word up with an answer, the remaining words had one fewer possible answers remaining. For example, a student might begin by matching *atmosphere* to "The air around the Earth," and then continue by matching *crater* to "Junk or pieces of rock" without considering the (already taken) answer "The air around the Earth." We wrote a computer program which enumerated all possible ways in which students might match up the five items with the five answers, assuming that they matched up items randomly without replacement. The resulting distribution of the number of correct answers is shown in Figure 6.1. The chance distribution of correct answers is not normally distributed, but fortunately analysis of variance is robust to departures from normality (personal communication, Brian Junker – Associate Professor of Statistics at Carnegie Mellon University). The median number correct for the chance distribution was 1. The mean number correct for the chance distribution was 119/120: (0 correct × 44 ways to get 0 correct) + (1 correct × 45 ways) + (2 correct × 20 ways) + (3 correct × 10 ways) + (4 correct × 1 way) = 119; 44 ways to get 0 correct + 45 + 20 + 10 + 1 = 120.

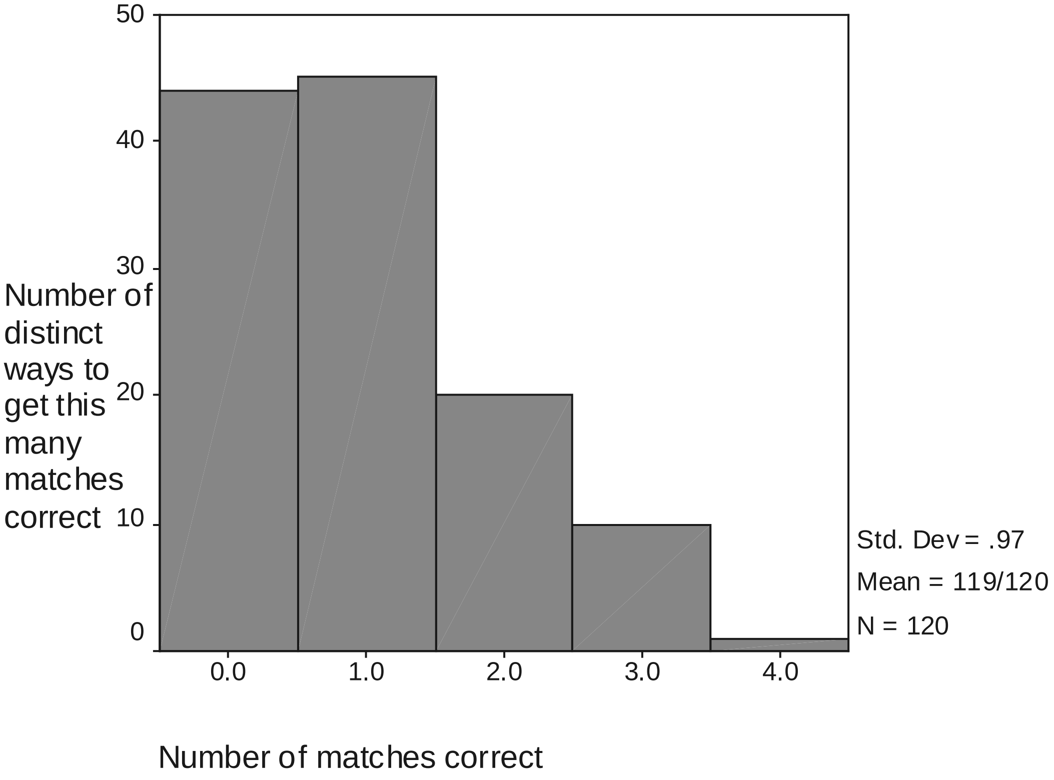


Figure 6.1. Chance distribution of number of correct answers on 5-item matching task. For this analysis we assumed that items were matched at random without replacement. Therefore, 5 items were scored as 4 correct because the fifth choice is completely determined by the previous four choices.

6.1.3 Experiment design

Did definitions help? We conducted an experiment comparing definitions to nonsemantic assistance, counterbalancing order of topic and assignment of topic to condition. Thus each subject saw both passages – one in each condition. Students read the passages on paper, with a paper test stapled to the passages and thus taken the same day: first one passage, then the test for that passage, then the other passage, then the test for the second passage. The design of this experiment is shown as a flowchart in Figure 6.2.

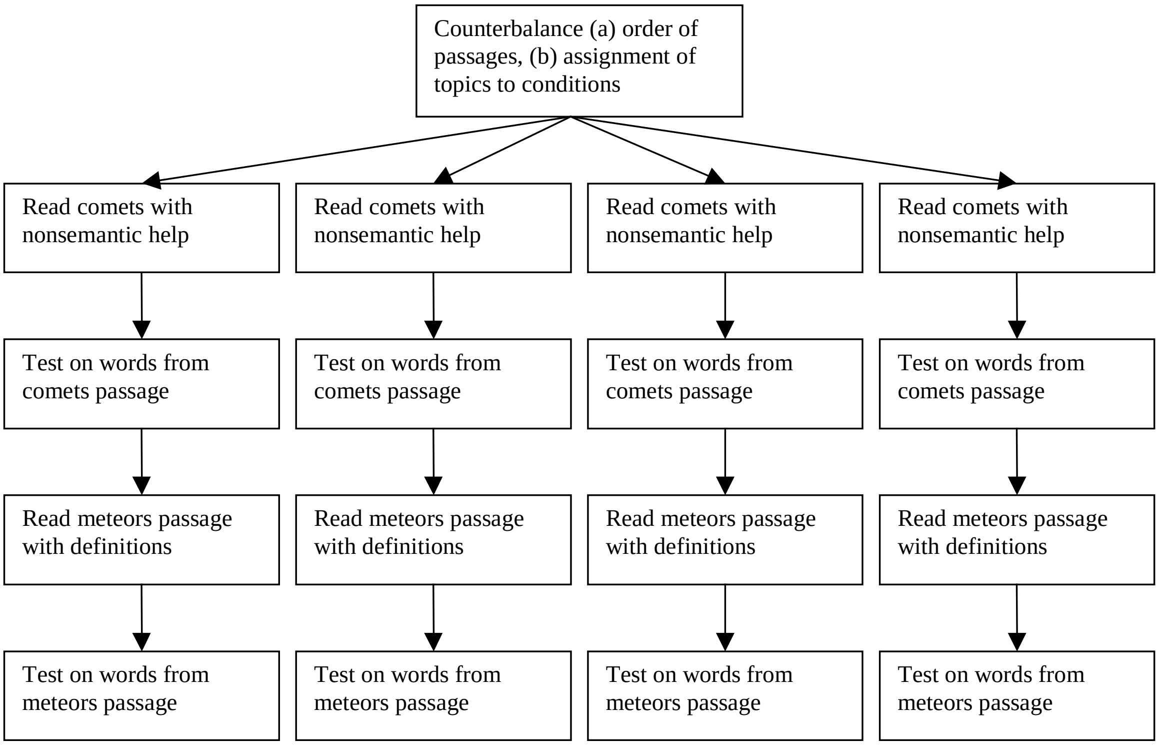

Counterbalance (a) order of passages, (b) assignment of topics to conditions
Read comets with nonsemantic help
Read comets with nonsemantic help
Read comets with nonsemantic help
Read comets with nonsemantic help
Test on words from comets passage
Test on words from comets passage
Test on words from comets passage
Test on words from comets passage
Read meteors passage with definitions
Read meteors passage with definitions
Read meteors passage with definitions
Read meteors passage with definitions
Test on words from meteors passage
Test on words from meteors passage
Test on words from meteors passage
Test on words from meteors passage

Figure 6.2. Flowchart for comets & meteors experiment.

We counterbalanced the order of topics (comets first, or meteors first), and the assignment of topics to conditions (comets to nonsemantic help and meteors to definitions, or vice versa). The tests contained one sample item (*ice* and *iron*, respectively) and five actual items. 41 students who had just finished 2nd through 5th grade participated, all from a low-income urban elementary school (Fort Pitt Elementary School).

6.1.4 Results

Explanations held an advantage over nonsemantic help. Analysis of variance (ANOVA) including a term for age showed a significant effect of definition on the matching task ($p = .041$). A t-test paired by student to compare the same student's responses in different categories showed that the definition helped: for the text that included definitions, students averaged 2.5 items right vs. 1.8 items right for the text plus nonsemantic help (p = .007). Thus students were able to make use of the information in a definition above and beyond the simple effect of an additional exposure.

6.1.5 Lessons learned from comets and meteors experiment

The comets and meteors experiment tested a same-day effect, and augmented text with either definitions or nonsemantic assistance. Our next experiment tested for effects on a later day, and compared definitions and natural contexts.

6.2 Can explanations add to natural contexts?: The limericks experiment

We wanted to explore the relative effectiveness of definitions and natural contexts. In Summer 2000, we conducted a within-subject experiment to explore the effectiveness of four conditions in which students might encounter an unknown word: not at all, in a definition alone, in a story alone, and in a story with an accompanying definition. This study was conducted during a month-long reading and math clinic at a low-income urban elementary school in Pittsburgh, Pennsylvania (Fort Pitt Elementary). During the clinic, each student was scheduled to spend 30 minutes per day on the Reading Tutor, Monday through Friday. School personnel, volunteers, and Project LISTEN staff provided supervision.

To minimize the impact of between-student variance, we designed this experiment within-subject. The materials were children’s limericks and experimenter-written explanations. We tested word familiarity and word knowledge: have you seen this word before, and do you know what it means. We designed the word familiarity question to provide a more basic measure of a student's knowledge of a word – easier than answering the multiple-choice questions that we used to measure word knowledge. We also expected to find main effects for both limerick and explanation, and had no strong prior expectations as to whether seeing an explanation alone would be better than seeing a limerick alone (Table 6.2).

	no limerick		limerick
no explanation		<	
	^		^
explanation		<	

Table 6.1. Expected direction of effects for word familiarity and word knowledge. A < B means that we expected cell B to have higher word familiarity and word knowledge than cell A.

In all, 29 students who had just finished 2nd - 5th grades completed the experiment, for a total of 232 trials, 58 trials for each of 4 conditions. Nine students had just finished second grade; nine, third grade; seven, fourth grade, and four had just finished fifth grade.

6.2.1 Texts

The stories were children's limericks by Edward Lear (19th cent.). There were eight limericks, with one target word each. The words were *dolorous*, *laconic*, *imprudent*, *innocuous*, *mendacious*, *oracular*, *irascible*, and *vexatious*.[5] The texts controlled for many factors:

1. Genre – all the limericks were poems.
2. Author – all the limericks were written by Edward Lear.
3. Intended audience – all the limericks were written for children.
4. Syntax – all of the limericks contained the target word in the last line, as follows: "That

[5] The test used the alternate spelling *vexacious.*

target word (old) *Person* of *Place*."

5. Word frequency – all of the target words occurred zero or one time in the Brown corpus (Kucera and Francis 1967), a million-word representative sampling of written English fiction and nonfiction in a variety of styles and domains – Francis and Kucera (1971 provides details. The words *mendacious* and *vexatious* occurred once; the other target words did not occur.
6. Part of speech – all target words were adjectives.
7. General semantic class – all target words described human personality traits.

Here is an example of a limerick:

There was an Old Man of Cape Horn,
Who wished he had never been born;
So he sat on a chair,
Till he died of despair,
That dolorous Man of Cape Horn.

The limericks are given in full in Appendix E.

We wrote the definitions for the target words to be as syntactically similar as possible. Each definition explained the words in ordinary language, following the advice given in McKeown (1993). For example: "We can say someone is dolorous if they are mournful, or feel really bad." The definitions are given in full in Appendix E.

6.2.2 Tests

We gave students a questionnaire on paper – for simplicity – with two questions for each target word (Appendix E). To exclude word recency effects (cf. Chapter 4), the answers on the questionnaire were different from the words that had appeared in the limericks and in the explanations. We administered the questionnaire one day after the student read the passages (25 out of 29 students), or two days later for students who were absent on the day after they read the passages (4 out of 29 students). The first question measured the student’s self-reported familiarity with the word; for example, “Have you ever seen the word *dolorous* before?”. Assuming none of the students had seen any of these words prior to the experiment, the “correct” answer would be “Yes” for the 75% of the words the student saw in the experiment, and “No” for the 25% of the words which that student did not see in the experiment. The second question measured the student’s ability to pick the correct meaning of the word from a set of four options. For example, “If someone is *dolorous* they must be… angry; sad; tired; afraid.” The second question had a chance performance level of 1 out of 4, or 25%.

Experiment design. There were four conditions (Table 6.2):

1. no limerick, no explanation;
2. no limerick, explanation;
3. limerick, no explanation;
4. limerick, explanation.

To minimize variance due to first- or last-item effects, we adopted a technique from the vocabulary assistance literature and held constant the order of presentation of the limericks. (Thus word and order were confounded together, but we were not interested in word or order effects.) Each condition occurred on two different words for each student. Thus, each student saw 6 of 8 words in the experiment – and probably had not seen those words prior to the experiment – with 2 words left as a no-exposure control. The assignment of words to conditions was set for a particular Reading Tutor computer.

Students read the texts with Reading Tutor assistance, seated at 12 different computers at tables as shown in Figure 6.3. Each computer was used by one student at a time. There were five sessions per day, with different students coming in for each session; the number of Reading Tutors in use varied from 5 to 12.

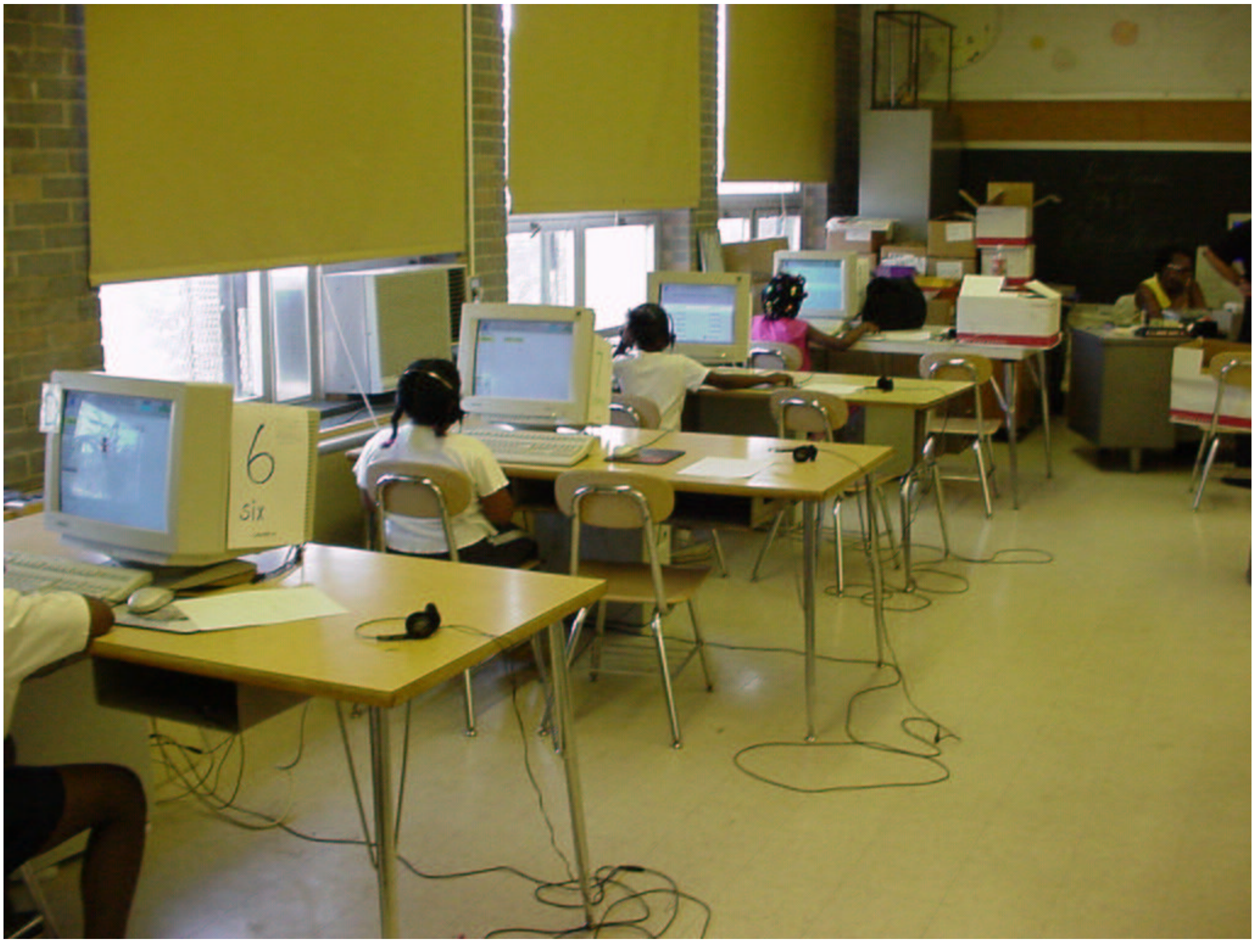


Figure 6.3. Children reading with the Reading Tutor at summer 2000 reading clinic. Photo credit: Mary Beth Sklar, Project LISTEN Educational Research Field Coordinator.

Figure 6.4 shows a flowchart of the design of the vocabulary limerick experiment.

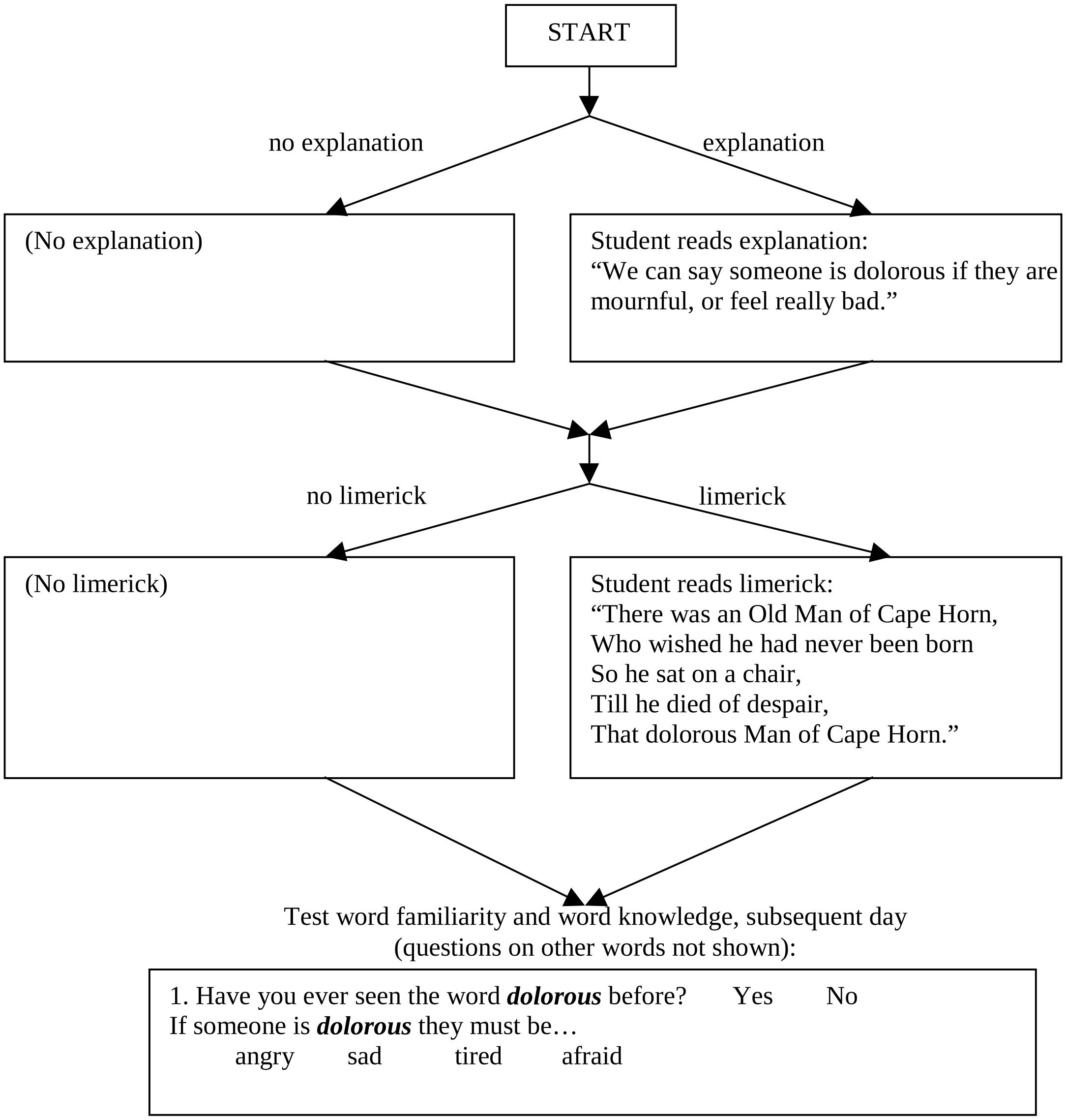

START
no explanation
explanation
(No explanation)
Student reads explanation:
“We can say someone is dolorous if they are mournful, or feel really bad.”
no limerick
limerick
(No limerick)
Student reads limerick:
“There was an Old Man of Cape Horn,
Who wished he had never been born
So he sat on a chair,
Till he died of despair,
That dolorous Man of Cape Horn.”
Test word familiarity and word knowledge, subsequent day
(questions on other words not shown):
1. Have you ever seen the word dolorous before? Yes No
If someone is dolorous they must be…
angry sad tired afraid

Figure 6.4. Limericks flowchart, showing one (of eight) limericks.

6.2.3 Treatment fidelity: 3% of trials affected by bug

We experienced a minor problem with treatment fidelity. A bug in the July 2000 version of the Reading Tutor used during this experiment allowed students to skip a sentence if they repeatedly clicked on the *Go* arrow (Figure 2.3, Chapter 2) when the Reading Tutor was already preparing to move on to the next sentence. As a result, some of the trials in which the student was supposed to see a sentence containing the target word – either a definition or a sentence from the original limerick – resulted in the student skipping the sentence.

To determine exactly how many trials the bug affected, we checked to see whether the sentence containing the target word had a corresponding student utterance recorded where the Reading Tutor heard the student say at least one word. (The Reading Tutor recorded every attempt at reading the sentence – see Chapter 2 for details.) Seven trials out of 232 had no nonempty student utterances for the sentence in which the target word appeared. Table 6.2 lists each of these trials affected by the bug.

ID	Grade	Word	Intended to display limerick?	Reading Tutor heard something for limerick sentence with target word?	Intended to display definition?	Reading Tutor heard something for definition?	*Yes* or *No* to "Have you seen this word?"	Multiple choice question …
CW	4	dolorous	Yes	No	No	No	Yes	Wrong
JR	2	vexatious	No	No	Yes	No	Not answered – coded as No	Right
MT	3	innocuous	Yes	Yes	Yes	No	Yes	Wrong
SJ	5	mendacious	No	No	Yes	No	No	Wrong
DG	2	imprudent	No	No	Yes	No	Yes	Wrong
DG	2	oracular	Yes	No	No	No	No	Wrong
DD	2	imprudent	Yes	No	Yes	Yes	Yes	Wrong

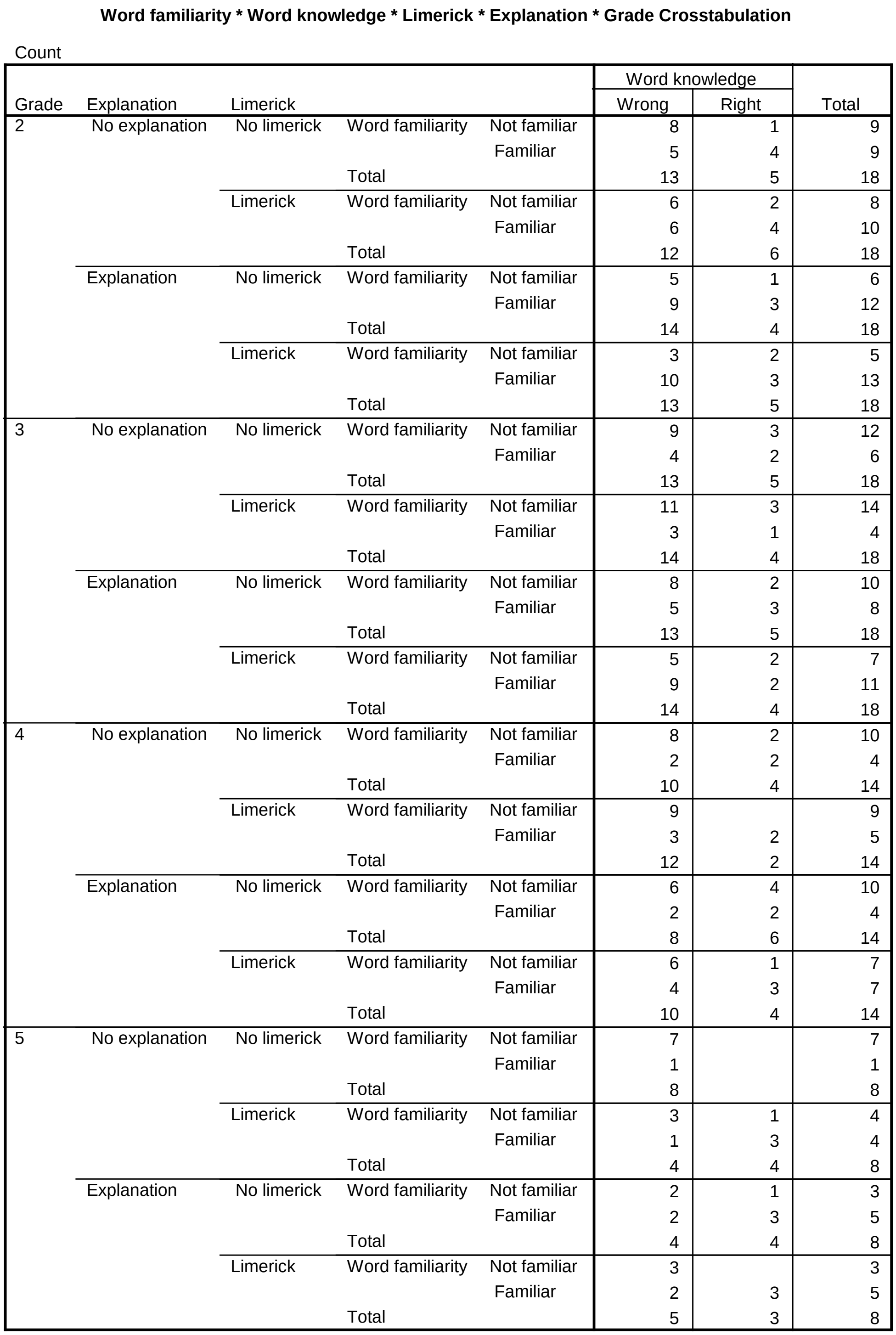
Table 6.2. Trials affected by sentence-skipping bug in Summer 2000 limericks experiment.

How should we handle these trials? One possibility is to "recode" the data – treat the trials that the bug affected as actually occurring in another condition. Such recoding confuses the data – the actual randomized decision was which sentence(s) to show, and the bug was a subsequent nonrandom effect. We could delete all trials for every student with any trial affected by the bug – but the bug affected 6 of the 29 students, so deleting the students would greatly shrink the subject pool. Finally, we could delete each trial that was affected by the bug. What was appropriate?

We wanted to answer two questions using data from the limericks experiment. First, what actually happened in the experiment? That is, how did students do on the words in various conditions – bug and all? To analyze the experiment as performed, we included all the data in the analysis. Second, what can we say about what we expect to happen if we adopt the vocabulary assistance policy represented by the experiment? That is, what can we say about future performance given that the bug has now been fixed? To predict future performance, we deleted the seven trials affected by the bug and re-ran the analysis. As it turned out, including or excluding the deleted trials did not substantially affect the results; so that the reader can see this for him- or herself, we report results both ways.

6.2.4 Results

Table 6.3 shows the results for the limericks experiment, crosstabulated by grade using SPSS. Table 6.4 shows the same results transformed into a more readable format. The two tables present complementary views of the data: Table 6.3 presents the raw data with subtotals, and Table 6.4 shows marginal probabilities. In two cases students did not answer one of the word familiarity questions; because the student failed say “Yes”, both such cases are coded the same as answering “No.” We discuss results as follows: word familiarity, word knowledge, and the relationship between familiarity and knowledge. We use a combination of (definitive) statistical comparisons and (exploratory) informal comparisons – noting the use of each along the way where appropriate.



Word familiarity * Word knowledge * Limerick * Explanation * Grade Crosstabulation

Count

					Word knowledge		
Grade	Explanation	Limerick			Wrong	Right	Total
2	No explanation	No limerick	Word familiarity	Not familiar	8	1	9
				Familiar	5	4	9
			Total		13	5	18
		Limerick	Word familiarity	Not familiar	6	2	8
				Familiar	6	4	10
			Total		12	6	18
	Explanation	No limerick	Word familiarity	Not familiar	5	1	6
				Familiar	9	3	12
			Total		14	4	18
		Limerick	Word familiarity	Not familiar	3	2	5
				Familiar	10	3	13
			Total		13	5	18
3	No explanation	No limerick	Word familiarity	Not familiar	9	3	12
				Familiar	4	2	6
			Total		13	5	18
		Limerick	Word familiarity	Not familiar	11	3	14
				Familiar	3	1	4
			Total		14	4	18
	Explanation	No limerick	Word familiarity	Not familiar	8	2	10
				Familiar	5	3	8
			Total		13	5	18
		Limerick	Word familiarity	Not familiar	5	2	7
				Familiar	9	2	11
			Total		14	4	18
4	No explanation	No limerick	Word familiarity	Not familiar	8	2	10
				Familiar	2	2	4
			Total		10	4	14
		Limerick	Word familiarity	Not familiar	9		9
				Familiar	3	2	5
			Total		12	2	14
	Explanation	No limerick	Word familiarity	Not familiar	6	4	10
				Familiar	2	2	4
			Total		8	6	14
		Limerick	Word familiarity	Not familiar	6	1	7
				Familiar	4	3	7
			Total		10	4	14
5	No explanation	No limerick	Word familiarity	Not familiar	7		7
				Familiar	1		1
			Total		8		8
		Limerick	Word familiarity	Not familiar	3	1	4
				Familiar	1	3	4
			Total		4	4	8
	Explanation	No limerick	Word familiarity	Not familiar	2	1	3
				Familiar	2	3	5
			Total		4	4	8
		Limerick	Word familiarity	Not familiar	3		3
				Familiar	2	3	5
			Total		5	3	8

Table 6.3. Results from Summer 2000 limericks study in SPSS crosstabulation. Grade here means grade just finished.

		no limerick	limerick
Grade 2	no explanation	*9 familiar/18 total (50%)* 1 right/9 unfamiliar (11%) 4 right/9 familiar (44%) **5 right/18 total (28%)**	*10 familiar/18 total (56%)* 2/8 unfamiliar (25%) 4/10 familiar (40%) **6 right/18 total (33%)**
	explanation	*12 familiar/18 total (67%)* 1 right/6 unfamiliar (17%) 3 right/12 familiar (25%) **4 right/18 total (22%)**	*13 familiar/18 total (72%)* 2 right/5 unfamiliar (40%) 3 right/13 familiar (23%) **5 right/18 total (28%)**
		no limerick	limerick
Grade 3	no explanation	*6 familiar/18 total (33%)* 3 right/12 unfamiliar (25%) 2 right/6 familiar (33%) **5 right/18 total (28%)**	*4 familiar/18 total (22%)* 3 right/14 unfamiliar (21%) 1 right/4 familiar (25%) **4 right/18 total (22%)**
	explanation	*8 familiar/18 total (44%)* 2 right/10 unfamiliar (20%) 3 right/8 familiar (38%) **5 right/18 total (28%)**	*11 familiar/18 total (61%)* 2 right/7 unfamiliar (29%) 2 right/11 familiar (18%) **4 right/18 total (22%)**
		no limerick	limerick
Grade 4	no explanation	*4 familiar/14 total (29%)* 2 right/10 unfamiliar (20%) 2 right/4 familiar (50%) **4 right/14 total (29%)**	*5 familiar/14 total (36%)* 0 right/9 unfamiliar (0%) 2 right/5 familiar (40%) **2 right/14 total (14%)**
	explanation	*4 familiar/14 total (29%)* 4 right/10 unfamiliar (40%) 2 right/4 familiar (50%) **6 right/14 total (43%)**	*7 familiar/14 total (50%)* 1 right/7 unfamiliar (14%) 3 right/7 familiar (43%) **4 right/14 total (29%)**
		no limerick	limerick
Grade 5	no explanation	*1 familiar/8 total (13%)* 0 right/7 unfamiliar (0%) 0 right/1 familiar (0%) **0 right/8 total (0%)**	*4 familiar/8 total (50%)* 1 right/4 unfamiliar (25%) 3 right/4 familiar (75%) **4 right/8 total (50%)**
	explanation	*5 familiar/8 total (63%)* 1 right/3 unfamiliar (33%) 3 right/5 familiar (60%) **4 right/8 total (50%)**	*5 familiar/8 total (63%)* 0 right/3 unfamiliar (33%) 3 right/5 familiar (60%) **3 right/8 total (38%)**
		no limerick	limerick
All grades	no explanation	*20 familiar/58 total (34%)* 6 right/38 unfamiliar (16%) 8 right/20 familiar (40%) **14 right/58 total (24%)**	*23 familiar/58 total (40%)* 6 right/35 unfamiliar (17%) 10 right/23 familiar (43%) **16 right/58 total (28%)**
	explanation	*29 familiar/58 total (50%)* 8 right/29 unfamiliar (28%) 11 right/29 familiar (38%) **19 right/58 total (33%)**	*36 familiar/58 total (62%)* 5 right/22 unfamiliar (31%) 11 right/36 familiar (31%) **16 right/58 total (28%)**

Table 6.4. Results for Summer 2000 limericks experiment, by grade (just finished) and overall.

6.2.4.1 Word familiarity results

Including all students and all trials, results were as follows (Table 6.5).

All students	no limerick	limerick
no explanation	20/58 familiar (34%)	23/58 familiar (40%)
explanation	29/58 familiar (50%)	36/58 familiar (62%)

Table 6.5. All students' performance on word familiarity question.

Explanations and familiarity. Explanations had a strong effect on self-reported familiarity, significant in a first-order logistic regression[6] model at $p < .001$: logistic regression coefficient 1.08 ± 0.32, with 99.9% confidence interval 0.02, 2.15.

Limericks and familiarity. Limericks exhibited a trend favoring a positive effect on familiarity, but not significantly: 0.50 ± 0.32, with 90% confidence interval -0.02, 1.03.

Notes on analysis. Excluding the trials affected by the sentence-skipping bug did not substantively change the results for main effect of explanation (1.08 ± 0.33) or main effect of limerick (0.50 ± 0.33). (Effect of explanation here means the overall effect of seeing an explanation, not the differential effectiveness of different explanations; likewise for limericks.) Including an interaction term in the model led to no significant interaction between explanation and limerick, and did not substantively change the results for main effect of explanation (1.26 ± 0.45) or main effect of limerick (0.67 ± 0.44).

Familiarity by grade. 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 2nd grade, 33% for 3rd grade, 29% for 4th grade, and 13% for fifth grade, on the 2 of 8 words they did not see in the study. Percentages reporting that they had seen the words before for the words they *did* see in the study reflects a propensity to answer Yes in general, as well as (presumably) increased awareness in older grades: 65% (35/54) for 2nd grade, 43% (23/54) for 3rd grade, 38% (16/42) for 4th grade, and 58% (14/24) for 5th grade.

Cell-by-cell comparisons. To characterize the data qualitatively, we looked at cell-by-cell comparisons. (We are not claiming statistical significance from such comparisons, merely seeking an overall sense of where the data lay.) The effect of explanation on familiarity was present across grade: In seven of eight cell-by-cell comparisons, the cell with explanations yielded higher familiarity than the cell without explanations. For example, Grade 2 no limerick/no explanation had 9 familiar out of 18 total, vs. Grade 2 limerick/no explanation which had 12 familiar out of 18 total. Likewise, the effect of limerick on familiarity was evident in seven out of eight cell-by-cell comparisons.

6.2.4.2 Word knowledge

The results for word knowledge are more nuanced than the results for familiarity.

Explanations and word knowledge. Considering all grades together, the trend favored a positive effect of explanations on word knowledge, but not significantly (0.24 ± 0.31).

Limericks and word knowledge. There was essentially no effect for seeing the target word in the limerick (-0.05 ± 0.31).

Notes on analysis. Excluding the trials affected by the sentence-skipping bug did not substantially change the results (explanation: 0.29 ± 0.32; limerick: -0.05 ± 0.32). Including an interaction term in the model resulted in no significant interaction between explanation and

[6] See glossary.

limerick (0.48 ± 0.63), no effect for explanation (0.00 ± 0.44), and a weak but surprisingly *negative* trend for limerick (-0.28 ± 0.43). What happened?

Consider the percentage of correct answers in the four conditions, aggregated for all 29 students (Table 6.6).

All students	no limerick	limerick
no explanation	14/58 right (24%)	16/58 right (28%)
explanation	19/58 right (33%)	16/58 right (28%)

Table 6.6. All students' performance on word knowledge question.

First, note that performance is close to chance (25%) in all the cells. Second, there is an apparent – and surprising – *decrease* in accuracy from the explanation/no limerick cell to the explanation/limerick cell. This apparent difference could of course simply be explained by random performance variation.

Word knowledge by grade. Now we consider students' performance on multiple-choice questions, disaggregated into individual grades. Second graders performed essentially at chance on multiple choice questions, getting 19 right out of 72 total, or 26%. Third graders also performed at chance: 18 right out of 72 total, or 25%. Furthermore, in seven out of eight cells for second grade and for third grade, the number correct was 4 out of 18 (22%) or 5 out of 18 (28%) – that is, as close to 25% as possible given that there were 18 trials in each cell.

But what about the fourth and fifth graders? Fourth graders got 16 right out of 56 total, or 29%. Fifth graders got 10 right out of 32 total, or 31%. In addition, fourth and fifth graders' performance was not as uniformly distributed across cells: performance ranged from 0 out of 8 (0%) in grade 5 with no explanation and no limerick to 4 out of 8 (50%) in grade 5 with explanation but no limerick and grade 5 with limerick but no explanation. What about cell-by-cell comparisons for fourth and fifth graders? Three out of four cell-by-cell comparisons for explanation vs. no explanation favor the explanation cell. However, only one out of four cell-by-cell comparisons of limerick vs. no limerick favors the limerick cell.

Because they performed similarly, we aggregated the fourth and fifth graders' data together (Table 6.7).

Fourth and fifth graders	no limerick	limerick
no explanation	18% right (4/22)	27% right (6/22)
explanation	45% right (10/22)	32% right (7/22)

Table 6.7. Fourth and fifth graders' data, aggregated together.

For fourth and fifth graders in a main-effects-only model, there was an effect favoring explanation, significant at p < .10: 0.89 ± 0.52, with 90% confidence interval 0.04, 1.74. There was no effect for limerick (-0.13 ± 0.51). Excluding trials affected by the sentence-skipping bug made no substantive difference (explanation: 0.88 ± 0.52; limerick: -0.12 ± 0.52). In a model including an interaction term for explanation and limerick, no effects were significant (explanation: 0.26 ± 0.72; limerick: -0.71 ± 0.69; explanation and limerick: 1.30 ± 1.04). Seeing an explanation alone was perhaps better than seeing the word in a limerick alone but again the trend was not significant (chi-square p = .310, McNemar's test p = .607). In sum, the evidence suggests that while second and third graders performed essentially at chance on the multiple-choice questions, fourth and fifth graders may have been helped by explanations of target words – but much less strikingly so than on the measure of familiarity.

6.2.4.3 The relationship between familiarity and word knowledge

There was a significant correlation (using the nonparametric correlation Kendall’s tau b) between self-reported familiarity and word knowledge, as shown in Table 6.8.

	grade 2	grade 3	grade 4	grade 5		overall
no explanation, no limerick	.372, p = .125	.088, p = .718	.300, p=.279	not applicable (answer is constant)		.269+, p = .055
explanation, no limerick	.094, p = .697	.194, p = .423	.091, p = .742	.258, p = .495		.110, p = .410
no explanation, limerick	.158, p = .514	.036, p = .883	.548*, p = .048	.500, p = .186		.288*, p = .031
explanation, limerick	-.169, p = .485	-.122, p = .615	.316, p = .254	.600, p =.112		.085, p = .526
all data	.113, p = .341	.049, p = .679	.271*, p = .044	.507**, p = .005		.187**, p=.004
all data except no exposure condition: explanation alone plus limerick alone plus explanation and limerick	.024, p = .863	.041, p = .771	.264+, p = .092	.438*, p = .032		.157*, p = .039

+ Significant at .10 level.
* Significant at .05 level.
** Significant at .01 level.

Table 6.8. Correlation between familiarity and word knowledge, by grade and condition.

Thus 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. Why? 2nd and 3rd graders performed at chance on the word knowledge test, so there was no correlation for them between word knowledge and familiarity – despite the increased reported familiarity after exposure to a word. Fourth and fifth graders, on the other hand, overall gained in familiarity; in addition, perhaps those who were able to extract some word meaning from the limericks remembered the word better than those who weren't.

Furthermore, consider how the correlation between familiarity and accuracy varied by grade and by condition (Table 6.8): an interesting pattern emerged. There is some evidence that the limerick strengthened the relationship between reported familiarity and actual word knowledge (correlation of 0.288, p = .031) – and no evidence that the definition did so (correlation of 0.110, p = .410) – even though the limerick was *less* effective overall. Thus, *if* students got anything out of the limerick, perhaps they had concentrated hard enough to remember it later – whereas the explanation did not require as much work from the reader, making the word not quite as easy to remember. This suggestion is preliminary and merely points the way to possible directions for future research. Also, we remind the reader that just because one correlation is significant and another is not does not necessarily imply that the first correlation is reliably greater than the second.

6.2.5 Lessons learned from limericks experiment

We now summarize lessons learned from the limericks experiment. Seeing an explanation helped all students in the study become more familiar with the word. Seeing the word in a limerick may have helped students become familiar with the word, but the trend was weak. In terms of learning the meaning of the word well enough to do better on a multiple-choice test, only explanations seemed to help – and only for fourth and fifth graders. We remind the reader that the exploratory nature of the preceding analysis – as opposed to a definitive analysis corrected for multiple comparisons – means that results derived from disaggregating the data by grade should be considered suggestive, not conclusive.

Furthermore, we note that reading the limerick might have had advantages beyond those revealed in the multiple-choice test – such as strengthening the relationship between familiarity and word knowledge.

Nonetheless, the effectiveness of the explanations and the lack of evidence for effectiveness of the limericks may be due to several factors. First, the limericks did contain information about the meaning of the target word, but required students to read and remember multiple sentences, make inferences, and remember their conclusions. The definitions, on the other hand, required students to read and remember only one sentence, and explicitly stated the meaning of the target word. Second, the genre of the limerick was poetry – to be read for understanding the poem and for appreciating the sound of the poem; the definitions were factual and clearly taught about the target word. Thus the text of the limerick did not imply that the task of the student was to learn the word, but the text of the definition definitely implied that the student was to learn the word. Finally, the target word was featured more prominently in the definition ("We can say someone is *dolorous*…") than in the limerick ("That *dolorous* Man of Cape Horn." as the last line of the poem, in passing.)

The results of the limericks study are not just same-day recency, since the test was given on a later day. Finally, the effect for word knowledge is due to remembering the meaning of the word – not just due to word recency from seeing the same word in the help and on the test, since the answers did not show up in the original text or in the definitions.

7 Conclusion

Reading is fundamental. Reading is comprehension: making meaning from print. Vocabulary underlies comprehension. We began with computer-assisted oral reading, and proceeded as follows. Improved story choice helped students encounter new material. Factoids comparing words in text to other words helped some students learn words. The Reading Tutor with Take Turns and factoids did better than a classroom control for 3rd graders on vocabulary learning – and even did comparably with one-on-one human tutoring. Finally, follow-on experiments pointed the way towards delivering improved vocabulary assistance.

To put these results in perspective, we note that the National Reading Panel observed that most vocabulary studies show effects only on experimenter-designed measures (NRP 2000) – not on standardized tests like the Woodcock Reading Mastery Test. Standard tests measure vocabulary so crudely that it is hard to achieve significant results when showing vocabulary growth, and even more so to show differences in growth between treatments. This dissertation not only achieved significant results on the Word Comprehension section of the Woodcock Reading Mastery Test (Chapter 5), but furthermore introduced and used two finer-grained techniques: first, new material read (Chapter 3); second, computer-constructed, in-context vocabulary questions (Chapter 4) as part of an embedded experiment (Mostow and Aist FF 2001; cf. Singh et al. 1999), encountered in the course of normal Reading Tutor use.

To achieve this goal, we built on a foundation of computer-assisted oral reading: Project LISTEN's Reading Tutor. Then, we developed, incorporated, and evaluated two improvements. First, we made the Reading Tutor take turns picking stories, which not only guaranteed that every student saw ~50% or more new material, but helped those students most who chose the fewest new stories themselves. (Such students were presumably those who needed the most practice in reading new text.) Second, we added automatically generated vocabulary assistance in the form of factoids – short comparisons to other words – and automatically generated vocabulary assessment in the form of multiple-choice questions. The factoids helped students answer multiple-choice questions – but only for third graders seeing rare words, and for single-sense rare words tested one or two days later. The multiple-choice questions explicitly operationalized Nagy et al.’s (1985) criteria for (Level 3) multiple choice questions, as we discussed in Section 4.3.1.2. Besides the factoids results, correlating the multiple choice questions with the Word Comprehension subtest of the Woodcock Reading Mastery Test demonstrated some validity. Finally, the examples of multiple-choice questions in Table 4.4 exposed additional constraints by violating them.

Follow-on experiments pointed the way towards even more effective vocabulary help, by presenting students with in-context explanations – and demonstrated students who had just finished 2nd through 5th grade gained word familiarity from exposure to words in the Reading Tutor, while 4th and 5th graders gained word knowledge from definitions as well.

Along the way, we used a variety of techniques, on timescales ranging from seconds to minutes to days to months (cf. Newell 1990’s time scale of human behavior). A story took seconds or minutes to choose, and minutes to read. We measured the effects of different story choice policies in the cumulative distribution of story choices over several months. Vocabulary assistance takes seconds to construct and present, and seconds to minutes to read. We measured the effects of vocabulary assistance either immediately (Chapter 6 comets & meteors experiment) or on a subsequent day (Chapter 4 factoids; Chapter 6 limericks). Finally, reading with the Reading Tutor took ~20 minutes/day for an entire school year – and we measured its effects during a year-long study.

Table 7.1 summarizes our experimental results, and Table 7.2 provides further detail on the vocabulary experiments. All three treatment conditions in the year-long study included a range of activities, including some directed at vocabulary development. Thus we compared three comprehensive treatments on a single aspect of learning to read, not three treatments aimed specifically at encouraging vocabulary development.

Goal	Chapter	Methodology	Key result
Improve story choice	Chapter 3	Modify Reading Tutor to take turns with the student at picking stories. Compare to Spring 1998 student-only story choice policy.	Higher percent of new material chosen in Fall 1999 (64.1%), vs. Spring 1998 (60.1%). Reading Tutor helped lower-performing students more.
Provide automatically generated vocabulary assistance	Chapter 4	Supplement stories with WordNet-extracted factoids; look for effect of factoids on answering multiple-choice questions. Compare trials with factoid + context to trials with context alone.	Factoids helped for the 189 trials with single-sense rare words tested one or two days later – significant at 95%, but exploratory.
Compare Reading Tutor to other reading instruction	Chapter 5	Analyze Word Comprehension portion of a larger Project LISTEN study comparing Reading Tutor with classroom instruction, one-on-one human tutoring	For third graders, Reading Tutor better than classroom control (effect size = 0.56, p = .042) and competitive with one-on-one human-assisted oral reading
Explore ways to improve vocabulary assistance	Chapter 6	Compare short explanations to nonsemantic assistance. Two texts with teacher-written definitions or nonsemantic assistance (COMET starts with C.)	At least when test is given in back of packet, students perform better on word-to-definition matching task when supplied with definitions (2.5 items right vs. 1.8).
	Chapter 6	Adapt limericks to vocabulary experiment. Compare no exposure vs. limerick alone vs. definition alone vs. limerick plus definition, all in Reading Tutor. Measure familiarity (“Have you seen this word before?”) and semantics (multiple-choice question on word meaning).	Strong effect of seeing explanations on familiarity. Trend favoring effect of seeing limericks on familiarity. Only 4th and 5th graders learned enough from definition to answer multiple-choice questions better.

Table 7.1. Summary of experimental results.

	Chapter 4: Factoids	Chapter 6: Comets and Meteors	Chapter 6: Limericks
Which students?	60 students in grades 2, 3 Centennial Elementary School Classroom setting	41 students who had just finished grades 2 through 5 Fort Pitt Elementary School Classroom setting	29 students in grades 2,3, 4, 5 Fort Pitt Elementary School Summer reading clinic setting
Which target words?	Words for which the Reading Tutor could automatically generate vocabulary assistance	Five domain-specific content words for each topic (10 words total)	Eight domain-independent but very rare adjectives
What kind of help?	Comparisons to other words, drawn from WordNet	Definitions written by story author (a teacher)	Experimenter-written context-specific explanations
When was help given?	Immediately before sentence containing target word	Immediately before sentence containing target words	Prior to limerick containing target word
At whose initiative?	Reading Tutor-selected using experimenter-written constraints	Teacher- (author-) selected words	Experimenter-selected words
What kind of text?	Stories already in the Reading Tutor	Two teacher-written nonfiction passages, one about comets and one about meteors	Eight children's limericks
Modality of text	Computer-assisted oral reading	Independent paper-based reading	Computer-assisted oral reading
Modality of vocabulary help	Help inserted in yellow pop-up boxes, to be read out loud in computer-assisted oral reading	Definitions inserted seamlessly into text, to be read independently on paper	Explanations inserted seamlessly into text, to be read out loud in computer-assisted oral reading
How tested?	Automatically generated multiple-choice questions, administered by the Reading Tutor	Five-item matching test, administered on paper, stapled to the text passages	4-item multiple-choice questions given on paper, subsequent day: Eight word familiarity yes-no questions and eight word knowledge
Results	Factoids helped for rare single-sense words tested one or two days later (44.1% correct with factoids vs. 25.8% correct without). Factoids also helped for third graders seeing rare words (42.0% with factoids vs. 36.2% without).	Definitions helped more than nonsemantic assistance on same-day matching task (2.5 items right vs. 1.8 items right.)	All students gained familiarity: 59/116 with limerick vs. 49/116 65/116 with definition vs. 43/116 Only 4th and 5th graders showed increased knowledge, and only for explanations: 13/22 right with limerick vs. 14/22 17/22 right with definition vs. 10/22

Table 7.2. Summary of vocabulary help experiments.

7.1 Contributions of this dissertation

This dissertation has contributed to a number of different fields, including intelligent tutoring systems, artificial intelligence and language technologies, and reading research. In particular, we have:

Designed mixed-initiative task choice for intelligent tutoring systems: Taking Turns. We have specified, implemented, and evaluated a mixed-initiative method for choosing stories – taking turns – that balances learner control and system control across time, and thus preserves some of the best of both absolutist policies. Taking turns may prove useful for choosing what task to work on next, in other tutoring systems.

Automated vocabulary assistance and assessment. We have also developed automatic generation of vocabulary assistance and assessment. We have not only explicated and operationalized construction of automatic assistance and assessment, but also identified limitations and thus additional requirements. The fact that automated assistance helped (sometimes) *despite* its flaws shows that this approach has potential. We have demonstrated that adding factoid vocabulary help to computer-assisted oral reading helped children learn vocabulary better than they would without factoids – but (so far) only for older students or single-sense rarer words.

Demonstrated the effectiveness of computer-assisted oral reading. We showed that computer-assisted oral reading – with Take Turns and factoids – can be more effective for third graders' vocabulary learning than a classroom control, and about as effective as human-assisted oral reading.

Identified a set of carefully controlled – yet authentically written – materials for vocabulary experiments. Carefully controlled vocabulary experiments sensitive enough to detect fine-grained learning require well-balanced texts. Materials originally written for some other purpose – not just to conduct a study – lend face validity to research studies. We identified and reported on a set of carefully controlled and freely available materials for use in future vocabulary studies: Edward Lear's limericks, together with our hand-written explanations and test items.

Explored which students can learn what aspects of vocabulary knowledge from what texts. Finally, we have helped clarify when students can make use of what kind of context for learning new words. Our limericks experiment showed that at least for older students, definitions increase word knowledge. However, story contexts may strengthen the relationship between familiarity and word knowledge.

7.2 Future work

We discuss future directions for story choice, and for giving help.

Story choice. The Reading Tutor story choice policy we have described took into account only the difficulty of a story, and not its content. Perhaps using story content might improve the Reading Tutor's choices, making them closer to a student's own choices in terms of acceptability while still making sure students read new and appropriately difficult material. For example, in order to allow the Reading Tutor to choose stories similar to those that an individual student chose, we could group the stories in the Reading Tutor into sets of twins by level and by topic, and have the Reading Tutor choose the twin of a story that the student chose.

Vocabulary assistance. We can consider the problem of vocabulary assistance as first choosing what words to give help on, and next choosing what kind of help to provide. In the factoids experiment, the Reading Tutor gave help on the words that it was feasible to give automated help on. In the comets & meteors experiment, we studied assistance for words that the stories' authors marked for explanation. In the limericks experiment, we studied assistance on general-purpose but extremely rare words – one or fewer occurrences per million words of English (Kucera and Francis 1967). While in the future we may consider revisiting the automated approach to providing vocabulary help, for now dividing labor between humans and computers to tap the strengths of both seems more promising. For example, a semi-automated approach might apply manual filtering to WordNet-extracted synonyms. In addition, selecting only (relatively) rare words to provide assistance on might improve the benefits of giving help.

7.3 Concluding remarks

We set out to demonstrate two claims, which we framed as improvements over factors in Equation 7.1:

$$\boxed{\frac{\text{New words learned}}{\text{Day}} = \frac{\text{Time on RT}}{\text{Day}} \times \frac{\text{Stories read}}{\text{Time on RT}} \times \frac{\text{New words seen}}{\text{Story read}} \times \frac{\text{New words learned}}{\text{New words seen}}}$$

Equation 7.1. New words learned per day on Reading Tutor.

First, by taking turns picking stories, an automated tutor that listens to children read aloud did indeed ensure that students read more new material than just their own choices would provide. In fact, students who chose the fewest new stories themselves benefited the most from the Reading Tutor's story choices – presumably such students needed the most practice reading new text. Second, by augmenting stories with semantic information about words, an automated reading tutor can help students learn words better than they would from the stories alone. Further experiments shed light on how to present effective vocabulary instruction, using short explanations of words. Finally, the 1999-2000 Reading Tutor with Take Turns and factoids outperformed a classroom control on Word Comprehension gains for third graders – and was even competitive with one-on-one human-assisted oral reading.

Glossary

boxplot

Boxplots were constructed (using SPSS) as follows. The heavy black line in the middle of the box is the median of the data. The bottom edge of the solid rectangle is just below the median of the bottom half of the data, also called the lower fourth. Likewise, the upper edge of the solid rectangle is just above the median of the upper half of the data, also called the upper fourth. The "T" bars represent data within 1.5 times the fourth spread of the data. (The fourth spread is the upper fourth minus the lower fourth.) Open circles (○) depict outliers farther away from the median than 1.5 times the fourth spread, but less than 3 times the fourth spread. Closed circles (●) – or in SPSS, stars (*) – depict data points farther away from the median than three times the fourth spread (Devore 1991 p. 28; SPSS 9.0). Numbering on closed circles is an (internal) case number used by SPSS.

logistic regression

A logistic regression model can be used to analyze an experiment where the outcome variable is binary – such as a yes or no answer, or a correct or incorrect answer – and the independent variables are categorical – such as grade. For this dissertation, we built logistic regression models by constructing general loglinear models in SPSS. In each case, the model was as follows:

$\log (p/1-p) = \alpha + \beta_b + \gamma_c + \delta_d + \ldots + \zeta_z$

where p equals the conditional probability of the desired outcome (*Yes* answer or correct answer) given each of the factors a through z.

SPSS estimates coefficients using maximum likelihood. The model reports coefficients that express the effect of changing one of the factors on the probability of the outcome, analogous to a linear regression model (y = a + bx) where changing b influences y.

Including a term for student made the analysis take into account student identity, analogous to McNemar’s test.

An overview and bibliography may be found at Garson, n.d. “Logistic Regression.” http://www2.chass.ncsu.edu/garson/pa765/logistic.htm. See also Menard (1995).

References

Aist, G. SR-CALL 1999. Speech recognition in computer assisted language learning. In K. C. Cameron (ed.), *Computer Assisted Language Learning (CALL): Media, Design, and Applications*. Lisse: Swets & Zeitlinger.

Aist, G. S. 1997. Challenges for a mixed initiative spoken dialog system for oral reading tutoring. AAAI 1997 Spring Symposium on Computational Models for Mixed Initiative Interaction. AAAI Technical Report SS-97-04.

Aist, G. and Mostow, J. ITS-PA 2001. Improving story choice in a reading tutor that listens. *Proceedings of the Fifth International Conference on Intelligent Tutoring Systems* (ITS'2000), p. 645. Montreal, Canada, June 2000. Poster Abstract. Available online as http://www.cs.cmu.edu/~aist/ITS2000-story-choice-abstract.doc.

Aist, G. S., and Mostow, J. CALL 1997. Adapting human tutorial interventions for a reading tutor that listens: using continuous speech recognition in interactive educational multimedia. In *Proceedings of CALL 97: Theory and Practice of Multimedia in Computer Assisted Language Learning*. Exeter, UK.

Aist, G. S., Mostow, J., Tobin, B., Burkhead, P., Corbett, A., Cuneo, A., Junker, B., and Sklar, M. B. AI-ED 2001. Computer-assisted oral reading helps third graders learn vocabulary better than a classroom control — about as well as human-assisted oral reading. *Proceedings of the Tenth Artificial Intelligence in Education (AI-ED) Conference*, San Antonio, Texas, May 2001.

American Guidance Service, n.d. *Bibliography for Woodcock Reading Mastery Tests – Revised (WRMT-R).* http://www.agsnet.com/Bibliography/WRMTRbio.html

The American Heritage® Dictionary of the English Language, Third Edition. 1996. Houghton Mifflin Company. Available online at dictionary.com.

Beck, Isabel, and Margaret McKeown. 1991. Conditions of vocabulary acquisition. In *Handbook of Reading Research* vol. 2: pp. 789-814. Mahwah, New Jersey: Lawrence Erlbaum.

Beck, I. L., McKeown, M. G., & McCaslin, E. S. 1983. Vocabulary development: All contexts are not created equal. *Elementary School Journal* 83: 177-181.

California State Board of Education. 1996. *Teaching Reading: A Balanced, Comprehensive Approach to Teaching Reading in Prekindergarten Through Grade Three*. http://www.cde.ca.gov/cilbranch/teachrd.htm. ISBN 0-8011-1276-1.

Carver, Ronald P. 1994. Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: Implications for instruction. *Journal of Reading Behavior* 26(4) pp. 413-437.

Devore, Jay L. 1991. *Probability and Statistics for Engineering and the Sciences*. 3rd edition. Pacific Grove, California: Brooks/Cole.

Donahue, P. L., Voelkl, K. E., Campbell, J. R., and Mazzeo, J. 1999. *NAEP 1998 Reading Report Card for the Nation and the States.* At http://nces.ed.gov/nationsreportcard/pubs/main1998/1999500.shtml. National Center for Education Statistics, Washington, DC.

Duke, N. K. 2000. Print environments and experiences offered to first-grade students in very low- and very high-SES school districts. *Reading Research Quarterly* 35(4): 456-457.

Edmark. 1997. Let’s Go Read. http://www.edmark.com/prod/lgr/island/.

Eller, Rebecca G., Pappas, Christine C., and Brown, Elga. 1988. The lexical development of kindergarteners: Learning from written context. *Journal of Reading Behavior* 20(1), pp. 5-24.

Fellbaum, Christiane, ed. 1998. *WordNet: An Electronic Lexical Database*. Cambridge MA: MIT Press.

Francis, W. N., and Kucera, H. 1971. *Brown Corpus Manual*. Providence, RI: Brown University. Downloaded from the World Wide Web on November 28, 2000 at http://visl.hum.ou.dk/itwebsite/corpora/corpman/BROWN/INDEX.HTM.

Garson, G. David. n.d. Logistic Regression. http://www2.chass.ncsu.edu/garson/pa765/logistic.htm.

Gipe, Joan P., and Richard D. Arnold. 1978. Teaching vocabulary through familiar associations and contexts. *Journal of Reading Behavior* 11(3): 281-285.

Hanna, Libby, Risden, Kirsten, Czerwinski, Mary, and Alexander, Kristin J. 1999. The role of usability research in designing children’s computer products. In *The Design of Children’s Technology*, Allison Druin (Ed.). San Fransisco: Morgan Kaufmann. pp. 3-26.

Huang, X. D., Alleva, F., Hon, H. W., Hwang, M. Y., Lee, K. F., and Rosenfeld, R. 1993. The Sphinx-II speech recognition system: An overview. *Computer Speech and Language* 7(2):137-148.

IBM. 1998. Watch Me Read. http://www.ibm.com/IBM/IBMGives/k12ed/watch.htm.

Juel, Connie. 1996. What makes literacy tutoring effective? *Reading Research Quarterly* 31(3), pp. 268-289.

Kucera, H. & Francis, W. N. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence, RI.

Kuhn, Melanie R., and Stahl, Steven A. 1998. Teaching children to learn word meanings from context: A synthesis and some questions. *Journal of Literacy Research* 30(1): 119-138.

Lear, Edward. 19th c. The Book of Nonsense. Available online from Project Gutenberg at ftp://sailor.gutenberg.org/pub/gutenberg/etext97/nnsns10.txt.

Madden, R., Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. 1973. *Stanford Achievement Test*. New York: Harcourt, Brace, Jovanovich, Inc. Cited in (McKeown 1985).

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

McKeown, Margaret G. 1985. The acquisition of word meaning from context by children of high and low ability. *Reading Research Quarterly* 20(4): 482-496.

McKeown, Margaret G., Isabel L. Beck, Richard C. Omanson, and Charles A. Perfetti. 1983. The effects of long-term vocabulary instruction on reading comprehension: A replication. *Journal of Reading Behavior* 15(1): 3-18.

Memory, David M. 1990. Teaching technical vocabulary: Before, during or after the reading assignment? *Journal of Reading Behavior* 22(1), pp. 39-53.

Menard, Scott. 1995. *Applied Logistic Regression Analysis.* vol. 106, Quantitative Applications in the Social Sciences. Sage Publications.

Merriam-Webster Student Dictionary, available online at wordcentral.com.

Mostow, J. 1996. A Reading Tutor that Listens (5-minute video). Presented at the DARPA CAETI Community Conference, November 19-22, 1996, Berkeley, CA.

Mostow, J., and Aist, G. FF 2001. Evaluating tutors that listen. In (K. Forbus and P. Feltovich, Eds.) *Smart Machines in Education: The coming revolution in educational technology*. MIT/AAAI Press. 2001.

Mostow, J., Aist, G., Burkhead, P., Corbett, A., Cuneo, A., Eitelman, S., Huang, C., Junker, B., Platz, C., Sklar, M. B., and Tobin, B. AI-ED poster 2001. A controlled evaluation of computer- versus human-assisted oral reading. Abstract of poster presented at the 10th International Conference on Artificial Intelligence in Education (AI-ED), San Antonio, Texas.

Mostow, J. and Aist, G. AAAI 1999. Authoring new material in a Reading Tutor that listens. *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, Orlando, FL, July 1999, pp. 918-919. In the refereed Intelligent Systems Demonstration track. Also presented at 37th Annual Meeting of the Association for Computational Linguistics (ACL’99), College Park, MD, June, 1999.

Mostow, J. and Aist, G. USPTO 1999. Reading and Pronunciation Tutor. United States Patent No. 5,920,838. Filed June 2, 1997; issued July 6, 1999. US Patent and Trademark Office.

Mostow, J., & Aist, G. CALICO 1999. Giving help and praise in a Reading Tutor with imperfect listening – Because automated speech recognition means never being able to say you're certain. *CALICO Journal* 16(3): 407-424. Special issue (M. Holland, Ed.), *Tutors that Listen: Speech recognition for Language Learning*.

Mostow, J., and Aist, G. S. AAAI 1997. The sounds of silence: towards automatic evaluation of student learning in a Reading Tutor that listens. In *Proceedings of the 1997 National Conference on Artificial Intelligence (AAAI 97*), pages 355-361.

Mostow, J., & Aist, G. PUI 1997. When speech input is not an afterthought: A Reading Tutor that listens. Workshop on Perceptual User Interfaces, Banff, Alberta, Canada, October 1997.

Mostow, J., Hauptmann, A. G., Chase, L. L., and Roth. S. 1993. Towards a Reading Coach that listens: Automatic detection of oral reading errors. In Proceedings of the Eleventh National *Conference on Artificial Intelligence (AAAI-93),* 392-397. Washington DC: American Association for Artificial Intelligence.

Mostow, J., Hauptmann, A., and Roth, S. F. 1995. Demonstration of a Reading Coach that listens. In *Proceedings of the Eighth Annual Symposium on User Interface Software and Technology*, Pittsburgh PA. Sponsored by ACM SIGGRAPH and SIGCHI in cooperation with SIGSOFT.

Mostow, J., Roth, S. F., Hauptmann, A. G., and Kane, M. 1994. A prototype Reading Coach that listens. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94),* Seattle WA. Selected as the AAAI-94 Outstanding Paper.

Nagy, William E., Herman, Patricia A., and Anderson, Richard C. 1985. Learning words from context. *Reading Research Quarterly* 20(2): 233-253.

National Reading Panel. 2000. *Teaching Children to Read.* http://www.nichd.nih.gov/publications/nrppubskey.cfm

Newell, A. 1990. *Unified Theories of Cognition*. Cambridge MA: Harvard UP.

Pinnell, G.S., Pikulski, J.J, Wixson, K.K., Campbell, J.R., Gough, P.B., & Beatty, A.S. 1995. *Listening to children read aloud.* Washington, DC: Office of Educational Research and Improvement. U.S. Department of Education.

Reinking, David, and Rickman, Sharon Salmon. 1990. The effects of computer-mediated texts on the vocabulary learning and comprehension of intermediate-grade learners. *Journal of Reading Behavior* 22(4), pp. 395-411.

Russell, M., Brown, C., Skilling, A., Series, R., Wallace, J., Bohnam, B., and Barker, P. 1996. Applications of Automatic Speech Recognition to Speech and Language Development in Young Children. *In Proceedings of the Fourth International Conference on Spoken Language Processing*, Philadelphia PA.

Schacter, John. 1999. *Reading Programs that Work: A Review of Programs from Pre-Kindergarten to 4th Grade*. Milken Family Foundation. PDF available from www.mff.org.

Scott, Judith A., and Nagy, William E. 1997. Understanding the definitions of unfamiliar verbs. *Reading Research Quarterly* 32(2), pp. 184-200.

Schwanenflugel, Paula J., Steven A. Stahl, and Elisabeth L. McFalls. 1997. Partial word knowledge and vocabulary growth during reading comprehension. *Journal of Literacy Research* 29(4): 531-553.

Shefelbine, John L. 1990. Student factors related to variability in learning word meanings from context. *Journal of Reading Behavior* 22(1): 71-97.

Singh, S., Kearns, M. S., Litman, D. J., & Walker, M. A. 1999. Reinforcement learning for spoken dialogue systems. In Proceedings of NIPS*99, to appear as S. A. Solla, T. K. Leen, and K.-R. Müller, (Editors), *Advances in Neural Information Processing Systems* 12. Cambridge, MA: MIT Press.

Snow, Catherine E., Burns, M. Susan, and Griffin, Peg, Eds. 1998. *Preventing Reading Difficulties in Young Children*. Washington D.C.: National Academy Press.

SPSS 9.0. 1998. Statistical software package. www.spss.com.

Wasik, B. A., and R. E. Slavin. 1993. Preventing early reading failure with one-to-one tutoring: A review of five programs. *Reading Research Quarterly* 28(2), 178-200.

Whines, N. 1999. Unpublished master's thesis, by Nick Whines, for Master of Arts in Design for Interactive Media, Middlesex University, London.

Williams, S.M., Nix, D., & Fairweather, P. 2000. Using Speech Recognition Technology to Enhance Literacy Instruction for Emerging Readers. In B. Fishman & S. O'Connor-Divelbiss (Eds.), *Proceedings of the Fourth International Conference of the Learning Sciences* (pp. 115-120). Mahwah, NJ: Erlbaum. http://www.umich.edu/~icls/proceedings/pdf/Williams.pdf

Zechmeister, E. B., Chronis, A. M., Cull, W. L., D'Anna, C. A., and Healy, N. A. 1995. Growth of a functionally important lexicon. *Journal of Reading Behavior* 27(2), pp. 201-212.

Appendix A: Experiment summaries in National Reading Panel format

This appendix contains a summary of the story choice study, from Chapter 3, and descriptions of the following other studies:

1. Factoids, from Chapter 4;
2. Comparison of Reading Tutor to other reading instruction, Chapter 5;
3. Comets and meteors, from Chapter 6; and,
4. Limericks, from Chapter 6.

Story choice, from Chapter 3

Spring 1998 vs. Fall 1999 story choice comparison, summarized using National Reading Panel coding scheme (NRP 2000).

	Spring 1998	Fall 1999
States or countries represented in sample	Pittsburgh and surrounding communities in western Pennsylvania, USA	
Number of different schools represented in sample	1: Fort Pitt Elementary	1: Centennial Elementary
Number of different classrooms represented in sample	3	6
Number of participants	>100 altogether 72 in WRMT analysis 24 in story choice analysis	144 in WRMT analysis 60 in story choice analysis
Age	7, 8, 9	7-11
Grade	2, 4, 5	2, 3
Reading levels of participants (prereading, beginning, intermediate, advanced)	Beginning	Intermediate
Whether participants were drawn from urban, suburban, or rural settings	Urban	Urban
Pretests administered prior to treatment	Woodcock Reading Mastery Test (WRMT): word attack, word identification, and passage comprehension subtests Oral reading fluency	Woodcock Reading Mastery Test (WRMT): word attack, word identification, word comprehension, and passage comprehension subtests Oral reading fluency
Socioeconomic status (SES)	Low SES	Mixed. 67% received free lunch 6.7% received reduced lunch → 75% received free or reduced lunch
Ethnicity	Predominantly Black/African-American	Predominantly White/European-American: ~35% black and ~65% white. 2 students may have reported multiethnic background (Hispanic/African-American/Hawaiian)
Exceptional learning characteristics	Unknown	1 student with cerebral palsy 2 students with significant speech impairments
First language	All except one or two were native speakers of English	All native speakers of English
Explain any selection restrictions that were applied to limit the sample of participants	None	Bottom half of class (as determined by teacher) selected to participate
Concurrent reading instruction received in classroom	Other reading instruction	Other reading instruction
How was sample obtained?	Sample was obtained by comparing samples from two different studies, each examining effectiveness of the Reading Tutor vs. other reading instruction	
Attrition	72 started in larger study	144 started

Number of participants lost per group during the study Was attrition greater for some groups that others?	5 moved 4 unavailable → 63 overall 24 using Reading Tutor	12 moved 1 unavailable for post-test → 131 overall (2 unavailable for readministering of post-test – post-test readministered to some students due to initial error) 60 using Reading Tutor
Setting of the study	Classroom	Classroom
Design of study	Nonequivalent control group design: behavior of existing groups from prior studies compared, controlling for age to address nonequivalence	
Describe all treatment and control conditions; be sure to describe nature and components of reading instruction provided to control group	Student-only story choice 1997-1998 Reading Tutor	Take-turns 1999-2000 Reading Tutor
Explicit or implicit instruction?	The Reading Tutor provides help on oral reading, consisting of large amounts of implicit instruction by modeling fluent reading and reading individual words. By pointing out specific instances of letter-to-sound rules (*a* here makes the sound /a/), the Reading Tutor also provides explicit instruction at the grapheme-to-phoneme level.	
Difficulty level and nature of texts	Authentic text ranging in level from pre-primer through fifth grade and including a mix of fiction and non-fiction. Some decodable text included to scaffold learning decoding skills.	Authentic text ranging in level from pre-primer through fifth grade and including a mix of fiction and non-fiction. Short factoids inserted into text (see Chapter 4).
Duration of treatments	Nominally 20-25 minutes per session, 5 sessions per week, for entire spring Actual usage: ~13 minutes/session, 1 day in 4-8	Nominally 20 minutes per session, 5 sessions per week, for entire fall Actual usage close to nominal guidelines
Was fidelity in delivering treatment checked?	Weekly visits by Project LISTEN personnel	2-3x/week visits by Project LISTEN personnel
Properties of teachers/trainers		
Number of trainers who administered treatment	One computer per classroom in study	One computer per classroom in study
Computer/student ratio	1:8	1:10-12
Type of computers	IBM-compatible personal computers running Windows NT	IBM-compatible personal computers running Windows NT
Special qualifications	The Reading Tutor listens to children read aloud	
Length of training	Not applicable	
Source of training		
Assignment of trainers to groups		
Cost factors	Personal computer costs ~$2500; cost of software depends on accounting for research and development costs	
List and describe other nontreatment independent variables included in the analysis of effects	Story level Grade	Story level Grade
List processes that were taught during training and measured during and at the end of training	Not applicable for comparison of story choice behavior	
List names of reading outcomes measured	Investigator-constructed quantitative measure of how much new material students were seeing: Percent new sentences per sentence encountered. No reason to suspect low reliability.	

List time points when dependent measures were assessed	Percent new material calculated over duration of each semester	
Any reason to believe that treatment/control groups might not have been equivalent prior to treatments?	Yes; students were from different schools and of different grades and ages. These two groups were selected for comparison because of the similarity in supervision, common location in classrooms, and length.	
Were steps taken in statistical analyses to adjust for any lack of equivalence?	Yes; analysis of variance controlled for grade.	
Result: Average percent new sentences per sentence encountered	See Chapter 3 for story choice study; Chapter 4 for factoids; Chapter 5 for Reading Tutor vs. classroom instruction vs. one-on-one human-assisted oral reading	
Difference: treatment mean minus control mean		
Effect size	Not applicable (Measure is a process variable, not an educational outcome variable.)	
Summary statistics used to derive effect size		
Number of people providing effect size information	Entire sample	Entire sample
Length of time to code study	Uncertain	
Name of coder	Aist	

Factoids, from Chapter 4

The factoids study was conducted at Centennial Elementary, concurrent with the Fall 1999 Story choice experiment summarized on the previous pages.

Reading Tutor vs. classroom instruction vs. human-assisted oral reading, from Chapter 5

The yearlong comparison of the Reading Tutor to other reading instruction was conducted at Centennial Elementary, with classroom control and human tutors as the comparison conditions.

Comets and meteors, from Chapter 6

The comets and meteors study was conducted at Fort Pitt Elementary during the spring of 2000.

Limericks, from Chapter 6

The limericks study was conducted at Fort Pitt Elementary during a summer reading clinic in July 2000.

Appendix B: Student-written stories

Student-written stories from Fall 1999 at Centennial Elementary School.

Last names modified to protect students’ identities thus: “S.” or “S------”.

Reader gender and initials	Story Title	How many times this student read and/or worked on (typing or narrating) a story with this title
fAS	Alish	1
fAS	Alisha’s Spelling Word	1
fAS	Alisha S.	1
fAS	Alisha S.,	23
fAS	Alisha S------	3
fAS	Anthony J.,	1
fAS	Gabe P.,	2
fAS	Older peopccccccccccccccccccccccccccccccccccccccccccccccome to mi par	2
fAS	Amanda S.,	1
fBC	Brandi C.,	1
fBC	Brittany G.,	1
fBG	Brittany G.,	2
fCD	Alisha S.,	1
fCP	Carly P.,	5
fDR	Dabnddnielle R.,	2
fDR	Danielle R.,	32
fDR	DanielR.,	29
fDR	niellerogresggvvddcckjfdxfccdscngvbhvbvbvbfuttuhttttttrfhhrbvbbb	9
fDR	ZoOoorrghh;sj’xcbgnhhhaetsazh,./;l;po]=-0;’/.mk,;/..;;;;[[/	1
fJB	Alto G.,	1
fJB	Jasmine B.,	17
fJB	Mary Beth S.,	1
fJB	Nicole C.,	5
fJB	Tijuan P.,	3
fJP	__NULL__	1
fJP	Jenna P.,	5
fJP	i like scool so muth	1
fJP	i like scool so muth bvnjbfjhbbnfmdbbdghdbbcghffhbhbg	2
fJP	i love my mom and dad	1
fJP	i love my mom and dad long ago i had a frend	1
fJP	I LOVE YOU YOU LOVE ME WERE A HAPPY FAMALY WETH A GRAT BIG HUG	1
fJP	Jessica P.,	27
fJP	why is this conpeter not werceng	1
fKG	Kierra G.,	37
fMG	Michelle G.,	17

fNC	Alto G.,	1
fNC	Jasmine B.,	1
fNC	Jordan A.,	1
fNC	Kierra G.,	1
fNC	Nicole C.,	32
fNC	Samuel F.,	1
fQB	boys	2
fQB	boysywyy hjdhgjsjiijk jhdjjftytrwrrg rwqtttffdwfrtfg1	1
fQB	Danielle R.,	1
fQB	Marcus W.,	1
fQB	Matt H.,	1
fQB	Quanisha B.,	67
fQB	quanisha b--- quanishab--- quaniasha b---	1
fQB	Story One, by K eith B- - ----	1
fQB	Story One, by Quanisha B ----	1
fQB	Story Three, by Quanisha B--- love matthew you tfgvsqed matthew	1
fQB	Story Three, by Quanisha B--- love matthew you tfgvsqed matthew.	1
fQB	Timesha A.,	1
fSB	Charles R.,	1
fSB	Michelle G.,	1
fSB	Samantha B.,	3
fSK	Sara K.,	2
fSO	Shawna O.,	1
fSW	Alisha’s Spelling Word	1
fSW	Symone W.,	15
fSW	SYMONE W----	2
fTA	niellerogresggvvddcckjfdxfccdscngvbhvbvbvbfuttuhtttttrfhhrbvbbb	1
fTA	Time6V Y78912235678912345678 9;’/.]=-[’]\=-[;’.,/.	1
fTA	Timesha A.,	43
fTB	Tawnei B.,	3
fTH	Brittany G.,	1
fTH	I had a dog and her name is DUTCHIS but some times she is	1
fTH	I had a dog and her name is DUTCHIS but some times she is bad an	1
fTH	I have a dog	1
fTH	I have a friend and her name Amanda	1
fTH	I have a friend and her name is AMANDA	3
fTH	I have and her	1
fTH	I HAVE PARENTS AND THEY ARE NICE AND I AM THE ONLY	1
fTH	I HAVE PARENTS AND THEY ARE NICE AND I AM THE ONLY CHILD THEY LO	1
fTH	iiincftsamkdsej e,ewtsZmu8wafuwqirqsijsgwhw6d23gbytqgqqwgygbwy t	2
fTH	Lee D.,	4
fTH	MY	1
fTH	Tinne H.,	4
fTH	TINNE H------	1
mAG	Alto	1
mAG	Alto .	1

mAG	Alto G.,	24
mAG	Alto is good and some time he is bad.	1
mAG	Boo ha HAHAHAHAHAHEHEHE	1
mAG	Boo ha HAHAHAHAHAHEHEHE.	1
mAG	HAHAHAheheheheheheheheyhehehehehehehehehehe	22
mAG	Hehehehehehehehehahahahahahahahahahahahahaha	3
mAG	Hehehehehehehehehahahahahahahahahahahahahaha.	1
mAG	hgiughotuogutgfghofuhbubhjigjiiohughugihugfhiufgihigfuh igfh igh	2
mAG	HIGHOGG9KOPII POHOIGIUJOIUOIUOUJ G OUOIUIU OI U UOHUOI OUOHIUGIK	2
mAG	hoo are you you are dum	1
mAG	Jasmine B.,	1
mAG	Jim is dum but I am cool and hard	3
mAG	jjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjjj	2
mAG	uhhhoupudpuofpuguh	1
mAJ	alisha selling words	1
mAJ	Anthony J.,	2
mAJ	Older peopccccccccccccccccccccccccccccccccccccccccccccome to mi par	1
mBE	Lee D.,	1
mBR	Brandon R.,	10
mBR	Jared T.,	1
mCR	Charles R.,	5
mDH	Lee D.,	1
mDM	Brittany G.,	2
mDM	Darnell H.,	5
mDM	Derek M.,	11
mDM	Glenn B.,	4
mDM	I had a dog and her name is DUTCHIS but some times she is bad an	1
mDM	I have a friend and her name is AMANDA	1
mDM	Lee D.,	6
mDS	Alisha S.,	1
mDS	Amanda S.,	1
mDS	Donald S.,	2
mDS	Older people play games and work and drive people home	2
mGB	Darnell H.,	2
mGB	Glenn B.,	29
mGB	Lee D.,	1
mGB	Story Eight by Glenn B.	1
mGP	ALISHA S STORY SIX.	1
mGP	Alisha S.,	1
mGP	Amanda S.,	1
mGP	Do you know that y2k	1
mGP	Do you know that y2k is the man . And the rock . Edge can bet th	1
mGP	Donald S.,	1
mGP	Gabe P.,	16
mGP	HBNGBGNBGNSJS WHSNJB JG G G F D D DD D D D D D D D D D D D D D D	1
mGP	kkkkkkkkkkkkkkkkUY6HJUIU799-0 -- -==0=-== 9=00;;;;;;;;;;;;;;;;;;	3

mGP	Leon F.,	3
mGP	Older peopccccccccccccccccccccccccccccccccccccccccccome to mi par	16
mGP	Older people play games and work and drive people home	8
mGP	Older people play games and work and drive people home.	1
mGP	SYMONE W----.Miss.H------- is nice today	2
mJA	Alto	1
mJA	Boo ha HAHAHAHAHAHEHEHE	2
mJA	HAHAHAhehehehehehehehehheyhehehehehehehehehehe	2
mJA	Jordan A.,	6
mJC	Jared T.,	1
mJC	Josh uaC.,	1
mJC	Joshua C.,	6
mJC	Travis R.,	2
mJC	Tyler B.,	3
mJC	Boo ha HAHAHAHAHAHEHEHE	1
mJC	Justin C.,	3
mJP	Alisha’s Spelling Word	2
mJP	ALISHA S STORY SIX.	1
mJP	Alisha S.,	1
mJP	alisha selling words	1
mJP	I Like WWF And wcw and ecw	1
mJP	Older peopccccccccccccccccccccccccccccccccccccccccccome to mi par	1
mJP	SYMONE W----	1
mJT	Jared T.,	1
mKB	;;	1
mKB	[][][][][][][][][][][][][]	1
mKB	‘‘	1
mKB	1 2 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7b	1
mKB	1 2 3 4 5 6 7 8 9 0 2 3 4 5 6 7 8 9 0 3 4 5 6 7 8 9 0	1
mKB	1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9	2
mKB	1 3 2 4 76 778 Y6 78 57 5 6	1
mKB	1 3 4 2 5 4 8 0 2 6 5 7 4 2 6	1
mKB	1.800.9.3475	1
mKB	1== ===============	3
mKB	123456789 123456789 123456789 123456789	1
mKB	(lots of numbers)	1
mKB	(different numbers)	3
mKB	(more numbers)	2
mKB	9 - 1 1 9 - 1 1 9 - 1 1 - 9 1 1 9 - 1 1	1
mKB	9 -1 1	1
mKB	9 -1 1 9 - 1 1 9 - 1 1	1
mKB	abcdefghijklmnopqrstuvwxyz	1
mKB	as as as as as as as as as as as as as as as as as as as as as a	1
mKB	awawawaawwawawawawawawawawawawa	4
mKB	back go	1
mKB	boys	1

mKB	go backghdfggrg	1
mKB	gobackbackgo	1
mKB	HAPPY BIRTHDAY	1
mKB	HVETHGGHDFHJGHRJHGJFHGJUGHDGHHDHJYUFYEUGFY RUFHEUFHERUYFHEURYFERU	5
mKB	Keith B.,	11
mKB	keith keith keith keith keith keith keih keith	1
mKB	Mhaehdgffshcrhhgfdfhggdgehjdrhgjfgfufyuyurehgduygrfegrarcus W.,	1
mKB	mom mom mom	2
mKB	Mytfedsikutfxsgghyorpebzvaqfpbnvcfghwiukhjfvxzaggkkdx	1
mKB	october oct	1
mKB	QWERTWERTYUYUIOOP[[]ASDFGHJKL;’	1
mKB	QWQWQW QWQWQWQWQWQWWQWQWQWQWQWQWQWQWQWQW WQWWQWQWQWQWQWQW QW QW Q	4
mKB	RED	1
mKB	red red red red red red red red	1
mKB	spelling	1
mKB	Story One, by K eith B- - ----	1
mKB	Story One, by Quanisha B ----	1
mKB	Story One, by vbbdvhhchbgdfjkfhh K eith B - - ----	1
mKB	Story Three, by Quanisha B---- love matthew you tfgvsqed matthew	1
mKB	timesha jessica mattheu	1
mKB	trevor trevor marcus marcus	1
mKB	UABCDEFGHIJKLOPLOPPQADRGDTTTTTTTTTTTTDanielle R.,	1
mLD	Brittany G.,	1
mLD	Lee D.,	5
mLF	Alisha S.,	2
mLF	Amanda S.,	1
mLF	Anthony J.,	3
mLF	Donald S.,	1
mLF	Gabe P.,	4
mLF	I Like WWF And wcw and ecw	2
mLF	I Like WWF And wcw and ecw . spell BIKE AND THE ROCK AND STONE C	1
mLF	Jordan S.,	2
mLF	Leon F jOHN p------ dAYTONA jORDAN aLISHA	3
mLF	Leon F.,	33
mLF	leon I Like WWF And wcw and ecw	2
mLF	Lon F.,	3
mLF	Symone W.,	1
mLG	boys	1
mLG	Lucas G.,	2
mLG	Marcus W.,	1
mLG	Mytfedsikutfxsgghyorpebzvaqfpbnvcfghwiukhjfvxzaggkkdx	1
mLG	Story One, by K eith B- - ----	1
mMH	ghkhkkhykfjkggkhkkhhgtitttvcfdderrsawkkkkkkkiuuuiuuiuuuuiuiiujjh	1

mMH	Marcus W.,	1
mMH	Matt H.,	43
mMH	MFinive.catdoggoggykjiuggigigkyghhkgggjhjhjgjhgggg.att H.,	5
mMW	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 23 2	1
mMW	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 2	3
mMW	a b c d e f g h i j k l n o p q r s t u v w x y z	1
mMW	a b c d e f g h i j k l n o p q r s t u v w x y z yellow 0range	1
mMW	abcdefghijklmnopqrstuvwxyz	1
mMW	abcdefghijklmnopqrstyvyz..marcus..abc	13
mMW	D UOY	1
mMW	Marcus W.,	6
mMW	marcus w------- pokemo	2
mMW	Mhaehdgffshcrhhgfdfhggdgehjdrhgjfgfufyuyurehgduygrfegrarcus W.,	4
mMW	Mytfedsikutfxsgghyorpebzvaqf	1
mMW	Mytfedsikutfxsgghyorpebzvaqfpbnvcfghwiukhjfvxzaggkkdx	5
mMW	Story One, by Mcarus W,,tuhgtrfdc	1
mMW	yellow orange green rad	1
mSF	Samuel F.,	4
mSF	store.	1
mSK	Shan K.,	1
mTB	Travis R.,	1
mTB	Tyler B.,	8
mTP	Nicole C.,	1
mTP	Samuel F.,	1
mTP	Tijuan P.,	39
mTR	Travis R.,	5

Appendix C: Story choice data

Stories started and finished by who chose the story and by level.

New/Old? * Story level * Start/Finish * Who chose? Crosstabulation

Count

				Story level								Total
Who chose?	Start /Finish			K	A	B	C	D	E	H	U	
Reading Tutor	Start	New/Old?	New	912	1310	1842	645	309	344	41		5403
		Total		912	1310	1842	645	309	344	41		5403
	Finish	New/Old?	New	522	541	246	141	54	5			1509
		Total		522	541	246	141	54	5			1509
Student	Start	New/Old?	New	293	367	243	83	113	107	42	462	1710
			Old	1078	122	95	20	16	3	5	533	1872
		Total		1371	489	338	103	129	110	47	995	3582
	Finish	New/Old?	New	192	210	89	32	17		3	115	658
			Old	712	75	49	10	4	2		155	1007
		Total		904	285	138	42	21	2	3	270	1665

New/Old? * Story level * Start/Finish * Who chose? * ID Crosstabulation

Count

Count													
					Story level								Total
ID	Who chose?	Start /Finish			K	A	B	C	D	E	H	U	
fAE	Reading Tutor	Start	New/Old?	New	11	13	8	9	1	3	1		46
			Total		11	13	8	9	1	3	1		46
		Finish	New/Old?	New	9	5	2	1	1				18
			Total		9	5	2	1	1				18
	Student	Start	New/Old?	New	6								6
				Old	23								23
			Total		29								29
		Finish	New/Old?	New	5								5
				Old	19								19

			Total		24								24
fAM	Reading Tutor	Start	New/Old?	New	12	16	8	4	2	3	1		46
			Total		12	16	8	4	2	3	1		46
		Finish	New/Old?	New	10	10	1	1	1				23
			Total		10	10	1	1	1				23
	Student	Start	New/Old?	New	4	5							9
				Old	26	1							27
			Total		30	6							36
		Finish	New/Old?	New	4	4							8
				Old	19	1							20
			Total		23	5							28
fAS	Reading Tutor	Start	New/Old?	New	12	60	331	11		1	1		416
			Total		12	60	331	11		1	1		416
		Finish	New/Old?	New	11	15	5	1					32
			Total		11	15	5	1					32
	Student	Start	New/Old?	New	4	9	10	1	2	1	1	21	49
				Old	19	6	1					13	39
			Total		23	15	11	1	2	1	1	34	88
		Finish	New/Old?	New	4	9	3					3	19
				Old	14	3	1					5	23
			Total		18	12	4					8	42
fAS	Reading Tutor	Start	New/Old?	New	7	17	15	1	4	9			53
			Total		7	17	15	1	4	9			53
		Finish	New/Old?	New	7	9	7	1	1	1			26
			Total		7	9	7	1	1	1			26
	Student	Start	New/Old?	New	3	13	10		2	7		1	36
				Old	3	4	3		1	1			12
			Total		6	17	13		3	8		1	48
		Finish	New/Old?	New	3	7	7		2			1	20
				Old	3	3	2		1	1			10
			Total		6	10	9		3	1		1	30
fBC	Reading Tutor	Start	New/Old?	New	11	17	34	1	2	1	1		67
			Total		11	17	34	1	2	1	1		67
		Finish	New/Old?	New	11	10	9						30
			Total		11	10	9						30
	Student	Start	New/Old?	New	10	15	9	1			1	2	38
				Old	30	2	4						36
			Total		40	17	13	1			1	2	74
		Finish	New/Old?	New	4	9	3	1					17
				Old	13	2	2						17
			Total		17	11	5	1					34
fBG	Reading Tutor	Start	New/Old?	New	13	54	39	10	2	1	1		120
			Total		13	54	39	10	2	1	1		120
		Finish	New/Old?	New	10	11	7	4	1				33
			Total		10	11	7	4	1				33

	Student	Start	New/Old?	New	7	27	4	2				1	41
				Old	14	11	1					1	27
			Total		21	38	5	2				2	68
		Finish	New/Old?	New	5	6	1					1	13
				Old	12	1	1						14
			Total		17	7	2					1	27
fCD	Reading Tutor	Start	New/Old?	New	7	7	9		25	14	1		63
			Total		7	7	9		25	14	1		63
		Finish	New/Old?	New	6	4	2		8				20
			Total		6	4	2		8				20
	Student	Start	New/Old?	New	6	15	5	1	9	16		1	53
				Old	1	1			1				3
			Total		7	16	5	1	10	16		1	56
		Finish	New/Old?	New	5	6	4	1	1				17
				Old		1			1				2
			Total		5	7	4	1	2				19
fCP	Reading Tutor	Start	New/Old?	New	12	14	18				1		45
			Total		12	14	18				1		45
		Finish	New/Old?	New	10	11	7						28
			Total		10	11	7						28
	Student	Start	New/Old?	New	5	13	12					1	31
				Old	5	3	1					4	13
			Total		10	16	13					5	44
		Finish	New/Old?	New	5	7	4					1	17
				Old	4	2	1					2	9
			Total		9	9	5					3	26
fDL	Reading Tutor	Start	New/Old?	New	13	28	11	13	4	1	1		71
			Total		13	28	11	13	4	1	1		71
		Finish	New/Old?	New	11	20	5	6					42
			Total		11	20	5	6					42
	Student	Start	New/Old?	New	4	8	2				1		15
				Old	42	3							45
			Total		46	11	2				1		60
		Finish	New/Old?	New	4	6	2						12
				Old	31	3							34
			Total		35	9	2						46
fDR	Reading Tutor	Start	New/Old?	New	12	23	20	13	6	4			78
			Total		12	23	20	13	6	4			78
		Finish	New/Old?	New	11	11	4	2	2				30
			Total		11	11	4	2	2				30
	Student	Start	New/Old?	New	4	1					3	41	49
				Old	45						1	32	78
			Total		49	1					4	73	127
		Finish	New/Old?	New	3	1						2	6
				Old	30							4	34
			Total		33	1						6	40

fJB	Reading Tutor	Start	New/Old?	New		3	44	3	35	21	1		107
			Total			3	44	3	35	21	1		107
		Finish	New/Old?	New		3	8	1	4	1			17
			Total			3	8	1	4	1			17
	Student	Start	New/Old?	New	8	4	7				1	15	35
				Old	4	1	1					12	18
			Total		12	5	8				1	27	53
		Finish	New/Old?	New	6	4	1					4	15
				Old	3	1						5	9
			Total		9	5	1					9	24
fJE	Reading Tutor	Start	New/Old?	New	11		4	4			1		20
			Total		11		4	4			1		20
		Finish	New/Old?	New	9		2	1					12
			Total		9		2	1					12
	Student	Start	New/Old?	New	6		1						7
				Old	13								13
			Total		19		1						20
		Finish	New/Old?	New	5								5
				Old	9								9
			Total		14								14
fJP	Reading Tutor	Start	New/Old?	New	13	23	22	13	5	2			78
			Total		13	23	22	13	5	2			78
		Finish	New/Old?	New	10	16	5	4	1				36
			Total		10	16	5	4	1				36
	Student	Start	New/Old?	New	5							2	7
				Old	49							3	52
			Total		54							5	59
		Finish	New/Old?	New	4							1	5
				Old	41								41
			Total		45							1	46
fJP	Reading Tutor	Start	New/Old?	New	2	3	25	7	9	38			84
			Total		2	3	25	7	9	38			84
		Finish	New/Old?	New	2	3	13	1	1				20
			Total		2	3	13	1	1				20
	Student	Start	New/Old?	New	3	5	7	3		2		8	28
				Old		1	3					26	30
			Total		3	6	10	3		2		34	58
		Finish	New/Old?	New	3	4	2	1				5	15
				Old		1	3					8	12
			Total		3	5	5	1				13	27
fKG	Reading Tutor	Start	New/Old?	New	36	84	12				1		133
			Total		36	84	12				1		133
		Finish	New/Old?	New	9	21	2						32
			Total		9	21	2						32
	Student	Start	New/Old?	New	2	16		1				1	20

				Old	2	7						36	45
			Total		4	23		1				37	65
		Finish	New/Old?	New	1	8		1				1	11
				Old		6						6	12
			Total		1	14		1				7	23
fLG	Reading Tutor	Start	New/Old?	New	10	17	18	14	4	1			64
			Total		10	17	18	14	4	1			64
		Finish	New/Old?	New	10	13	10	8					41
			Total		10	13	10	8					41
	Student	Start	New/Old?	New	6	13	7	3	1		1		31
				Old	26	5	2						33
			Total		32	18	9	3	1		1		64
		Finish	New/Old?	New	5	10	2	3					20
				Old	25	5	1						31
			Total		30	15	3	3					51
fMG	Reading Tutor	Start	New/Old?	New	11	26	9	8	3	4			61
			Total		11	26	9	8	3	4			61
		Finish	New/Old?	New	10	14	5	3	2				34
			Total		10	14	5	3	2				34
	Student	Start	New/Old?	New	5						1	1	7
				Old	22							16	38
			Total		27						1	17	45
		Finish	New/Old?	New	4							1	5
				Old	16							8	24
			Total		20							9	29
fMW	Reading Tutor	Start	New/Old?	New		7					1		8
			Total			7					1		8
		Finish	New/Old?	New		5							5
			Total			5							5
	Student	Start	New/Old?	New		4							4
				Old		1							1
			Total			5							5
		Finish	New/Old?	New		2							2
			Total			2							2
fNC	Reading Tutor	Start	New/Old?	New		3	14	2	11	21			51
			Total			3	14	2	11	21			51
		Finish	New/Old?	New		2	9	1	4				16
			Total			2	9	1	4				16
	Student	Start	New/Old?	New		2	7	2	4	1		10	26
				Old			7					27	34
			Total			2	14	2	4	1		37	60
		Finish	New/Old?	New		2	5	2				2	11
				Old			4					5	9
			Total			2	9	2				7	20
fNC	Reading Tutor	Start	New/Old?	New	9	6	67	1	1		1		85

			Total		9	6	67	1	1		1		85
		Finish	New/Old?	New	8	5	4						17
			Total		8	5	4						17
	Student	Start	New/Old?	New	6	9	3						18
				Old	5	4	12						21
			Total		11	13	15						39
		Finish	New/Old?	New	6	5	2						13
				Old	4	1	5						10
			Total		10	6	7						23
fPO	Reading Tutor	Start	New/Old?	New	9	19	9	11	4	2	1		55
			Total		9	19	9	11	4	2	1		55
		Finish	New/Old?	New	9	12	7	8	1				37
			Total		9	12	7	8	1				37
	Student	Start	New/Old?	New	7	13	14	2	6	16			58
				Old	8	3	4	2					17
			Total		15	16	18	4	6	16			75
		Finish	New/Old?	New	6	9	5	1	1				22
				Old	5	3	3	2					13
			Total		11	12	8	3	1				35
fQB	Reading Tutor	Start	New/Old?	New	20	51	31	41	12	13			168
			Total		20	51	31	41	12	13			168
		Finish	New/Old?	New	12	13	2	4	2				33
			Total		12	13	2	4	2				33
	Student	Start	New/Old?	New	3	1	1	2	1		2	19	29
				Old	32							60	92
			Total		35	1	1	2	1		2	79	121
		Finish	New/Old?	New	2	1		2				5	10
				Old	18							12	30
			Total		20	1		2				17	40
fSB	Reading Tutor	Start	New/Old?	New	9	15	18	9	2	6	1		60
			Total		9	15	18	9	2	6	1		60
		Finish	New/Old?	New	8	3	2						13
			Total		8	3	2						13
	Student	Start	New/Old?	New	6							3	9
				Old	17							2	19
			Total		23							5	28
		Finish	New/Old?	New	6							1	7
				Old	11								11
			Total		17							1	18
fSK	Reading Tutor	Start	New/Old?	New	17	22	14	8	8	5			74
			Total		17	22	14	8	8	5			74
		Finish	New/Old?	New	11	14	5	2	1				33
			Total		11	14	5	2	1				33
	Student	Start	New/Old?	New	9					2		1	12
				Old	45							1	46
			Total		54					2		2	58

		Finish	New/Old?	New	4							1	5
				Old	35							1	36
			Total		39							2	41
fSO	Reading Tutor	Start	New/Old?	New	15	15	10	9	3		1		53
			Total		15	15	10	9	3		1		53
		Finish	New/Old?	New	12	12	4	3	1				32
			Total		12	12	4	3	1				32
	Student	Start	New/Old?	New	3	15	3	6	1			1	29
				Old	13	8	1						22
			Total		16	23	4	6	1			1	51
		Finish	New/Old?	New	3	9	1	2					15
				Old	12	4							16
			Total		15	13	1	2					31
fSW	Reading Tutor	Start	New/Old?	New	15	41	9	8	3	2	1		79
			Total		15	41	9	8	3	2	1		79
		Finish	New/Old?	New	12	29	2						43
			Total		12	29	2						43
	Student	Start	New/Old?	New	3	14	2	1	1	3		3	27
				Old	8	12						15	35
			Total		11	26	2	1	1	3		18	62
		Finish	New/Old?	New	3	11	1	1				3	19
				Old	6	10						10	26
			Total		9	21	1	1				13	45
fTA	Reading Tutor	Start	New/Old?	New	16	22	14	6	5	6			69
			Total		16	22	14	6	5	6			69
		Finish	New/Old?	New	10	9			2				21
			Total		10	9			2				21
	Student	Start	New/Old?	New	5	1		1	3	2	2	3	17
				Old	27				4			42	73
			Total		32	1		1	7	2	2	45	90
		Finish	New/Old?	New	4			1	1			1	7
				Old	14				1			1	16
			Total		18			1	2			2	23
fTB	Reading Tutor	Start	New/Old?	New	8	4	7	1	1	32	1		54
			Total		8	4	7	1	1	32	1		54
		Finish	New/Old?	New	8	3	3		1	1			16
			Total		8	3	3		1	1			16
	Student	Start	New/Old?	New	6	6	6	1		15		2	36
				Old	3							1	4
			Total		9	6	6	1		15		3	40
		Finish	New/Old?	New	6	2	4	1				1	14
				Old	3							1	4
			Total		9	2	4	1				2	18
fTH	Reading Tutor	Start	New/Old?	New	39	42	4	4		1	1		91
			Total		39	42	4	4		1	1		91

		Finish	New/Old?	New	13	3	2	1					19
			Total		13	3	2	1					19
	Student	Start	New/Old?	New	5	8						17	30
				Old	11							6	17
			Total		16	8						23	47
		Finish	New/Old?	New	2	1						9	12
				Old	4							1	5
			Total		6	1						10	17
mAG	Reading Tutor	Start	New/Old?	New		2	1	2	24	63	1		93
			Total			2	1	2	24	63	1		93
		Finish	New/Old?	New		1	1	1	5	1			9
			Total			1	1	1	5	1			9
	Student	Start	New/Old?	New		6		1	4	3	3	44	61
				Old		4			8		1	23	36
			Total			10		1	12	3	4	67	97
		Finish	New/Old?	New		2					1	4	7
				Old		4			1			6	11
			Total			6			1		1	10	18
mAJ	Reading Tutor	Start	New/Old?	New	20	7	3	9		2			41
			Total		20	7	3	9		2			41
		Finish	New/Old?	New	11	4	2	2					19
			Total		11	4	2	2					19
	Student	Start	New/Old?	New	9	3	2	3	15	3	1	3	39
				Old	18				2			1	21
			Total		27	3	2	3	17	3	1	4	60
		Finish	New/Old?	New	4	1	1	1	2			1	10
				Old	9							1	10
			Total		13	1	1	1	2			2	20
mBE	Reading Tutor	Start	New/Old?	New	13	14	168	2			1		198
			Total		13	14	168	2			1		198
		Finish	New/Old?	New	6	8	5	1					20
			Total		6	8	5	1					20
	Student	Start	New/Old?	New	12	9	35	1	1			1	59
				Old	12	4	5						21
			Total		24	13	40	1	1			1	80
		Finish	New/Old?	New	2	6	3		1			1	13
				Old	5	2	1						8
			Total		7	8	4		1			1	21
mBR	Reading Tutor	Start	New/Old?	New	52	17	10	7		1	2		89
			Total		52	17	10	7		1	2		89
		Finish	New/Old?	New	11	3	1						15
			Total		11	3	1						15
	Student	Start	New/Old?	New	15	5	9	3	2	7	1	3	45
				Old	29		1					8	38
			Total		44	5	10	3	2	7	1	11	83
		Finish	New/Old?	New	3	1	1					1	6

				Old	5							4	9
			Total		8	1	1					5	15
mCR	Reading Tutor	Start	New/Old?	New	10	31	16	12	5	3	1		78
			Total		10	31	16	12	5	3	1		78
		Finish	New/Old?	New	10	13	2	1					26
			Total		10	13	2	1					26
	Student	Start	New/Old?	New	6							1	7
				Old	32							4	36
			Total		38							5	43
		Finish	New/Old?	New	4							1	5
				Old	22							3	25
			Total		26							4	30
mDB	Reading Tutor	Start	New/Old?	New	12	73	16	17	5	2			125
			Total		12	73	16	17	5	2			125
		Finish	New/Old?	New	11	21	5	5	2				44
			Total		11	21	5	5	2				44
	Student	Start	New/Old?	New	5	6	2	1	2	2			18
				Old	45	6		1					52
			Total		50	12	2	2	2	2			70
		Finish	New/Old?	New	4	3							7
				Old	35	3		1					39
			Total		39	6		1					46
mDH	Reading Tutor	Start	New/Old?	New	14	3	1	2	21		1		42
			Total		14	3	1	2	21		1		42
		Finish	New/Old?	New	9	3		1	2				15
			Total		9	3		1	2				15
	Student	Start	New/Old?	New	9	4	14	3	21			1	52
				Old	10								10
			Total		19	4	14	3	21			1	62
		Finish	New/Old?	New	4	3	1	1	4				13
				Old	7								7
			Total		11	3	1	1	4				20
mDM	Reading Tutor	Start	New/Old?	New	20	37	306	10	6	2	1		382
			Total		20	37	306	10	6	2	1		382
		Finish	New/Old?	New	10	8	9	3	1				31
			Total		10	8	9	3	1				31
	Student	Start	New/Old?	New	9	6	15	1	1		3	17	52
				Old	20	2	19					13	54
			Total		29	8	34	1	1		3	30	106
		Finish	New/Old?	New	5	2	6					4	17
				Old	7	2	5					4	18
			Total		12	4	11					8	35
mDS	Reading Tutor	Start	New/Old?	New		3	188	71			1		263
			Total			3	188	71			1		263
		Finish	New/Old?	New		3	9	10					22

			Total			3	9	10					22
	Student	Start	New/Old?	New	1	6	7	5	2		2	6	29
				Old	3	1	19				3		26
			Total		4	7	26	5	2		5	6	55
		Finish	New/Old?	New	1	4	3	2			2	1	13
				Old	3		13						16
			Total		4	4	16	2			2	1	29
mGB	Reading Tutor	Start	New/Old?	New	14	53	26	40	8	5	1		147
			Total		14	53	26	40	8	5	1		147
		Finish	New/Old?	New	11	11	3	3					28
			Total		11	11	3	3					28
	Student	Start	New/Old?	New	8	10	1			1	2	18	40
				Old	30	1		1				15	47
			Total		38	11	1	1		1	2	33	87
		Finish	New/Old?	New	4	5						1	10
				Old	20			1				2	23
			Total		24	5		1				3	33
mGP	Reading Tutor	Start	New/Old?	New	28	28	24	22	9	4	1		116
			Total		28	28	24	22	9	4	1		116
		Finish	New/Old?	New	14	5	4	1					24
			Total		14	5	4	1					24
	Student	Start	New/Old?	New	1	6	2			3	4	41	57
				Old	18							15	33
			Total		19	6	2			3	4	56	90
		Finish	New/Old?	New	1	5	1					5	12
				Old	8							3	11
			Total		9	5	1					8	23
mJA	Reading Tutor	Start	New/Old?	New		4	6	35	4	4	1		54
			Total			4	6	35	4	4	1		54
		Finish	New/Old?	New		3	2	9	2				16
			Total			3	2	9	2				16
	Student	Start	New/Old?	New	2	5	5	5				6	23
				Old	1	1		4				5	11
			Total		3	6	5	9				11	34
		Finish	New/Old?	New	2	4	2	2				3	13
				Old	1	1		3				2	7
			Total		3	5	2	5				5	20
mJC	Reading Tutor	Start	New/Old?	New	108						1		109
			Total		108						1		109
		Finish	New/Old?	New	6								6
			Total		6								6
	Student	Start	New/Old?	New	8	5	1	3	12	1	1	8	39
				Old	1							5	6
			Total		9	5	1	3	12	1	1	13	45
		Finish	New/Old?	New								2	2
				Old								4	4

			Total									6	6
mJC	Reading Tutor	Start	New/Old?	New	8	4	15	4			1		32
			Total		8	4	15	4			1		32
		Finish	New/Old?	New	8	4	6	2					20
			Total		8	4	6	2					20
	Student	Start	New/Old?	New	5	2	9					2	18
				Old	3	3	7					2	15
			Total		8	5	16					4	33
		Finish	New/Old?	New	4	2	8					2	16
				Old	3	1	3						7
			Total		7	3	11					2	23
mJH	Reading Tutor	Start	New/Old?	New	3	19	8	2	1	14	1		48
			Total		3	19	8	2	1	14	1		48
		Finish	New/Old?	New	2	9	6	1	1	1			20
			Total		2	9	6	1	1	1			20
	Student	Start	New/Old?	New	2	10	7	1	1	7			28
				Old	1	2	2			2			7
			Total		3	12	9	1	1	9			35
		Finish	New/Old?	New	2	7	6	1	1				17
				Old	1	2	2			1			6
			Total		3	9	8	1	1	1			23
mJK	Reading Tutor	Start	New/Old?	New	10	8	12	8			1		39
			Total		10	8	12	8			1		39
		Finish	New/Old?	New	9	7	11	5					32
			Total		9	7	11	5					32
	Student	Start	New/Old?	New	7	6	2	1		4			20
				Old	20	3	1						24
			Total		27	9	3	1		4			44
		Finish	New/Old?	New	6	5							11
				Old	16	1	1						18
			Total		22	6	1						29
mJP	Reading Tutor	Start	New/Old?	New	28	21	18	10	4	3	1		85
			Total		28	21	18	10	4	3	1		85
		Finish	New/Old?	New	14	11	10	6					41
			Total		14	11	10	6					41
	Student	Start	New/Old?	New	3	7	9	2	5		2	7	35
				Old	26	1	1					1	29
			Total		29	8	10	2	5		2	8	64
		Finish	New/Old?	New	1	6	2	1				5	15
				Old	22	1	1					1	25
			Total		23	7	3	1				6	40
mJT	Reading Tutor	Start	New/Old?	New	13	1		2		1			17
			Total		13	1		2		1			17
		Finish	New/Old?	New	11	1							12
			Total		11	1							12

	Student	Start	New/Old?	New	3	1						1	5
				Old	10								10
			Total		13	1						1	15
		Finish	New/Old?	New	3	1						1	5
				Old	9								9
			Total		12	1						1	14
mKB	Reading Tutor	Start	New/Old?	New	12	45	36	39	9	8	1		150
			Total		12	45	36	39	9	8	1		150
		Finish	New/Old?	New	11	21	5	3					40
			Total		11	21	5	3					40
	Student	Start	New/Old?	New	4	2	1			2	2	60	71
				Old	31							11	42
			Total		35	2	1			2	2	71	113
		Finish	New/Old?	New	4	1						13	18
				Old	24							8	32
			Total		28	1						21	50
mLD	Reading Tutor	Start	New/Old?	New	37	7	2		1		1		48
			Total		37	7	2		1		1		48
		Finish	New/Old?	New	13	5							18
			Total		13	5							18
	Student	Start	New/Old?	New	4	1				2		3	10
				Old	23							3	26
			Total		27	1				2		6	36
		Finish	New/Old?	New	2							1	3
				Old	8								8
			Total		10							1	11
mLF	Reading Tutor	Start	New/Old?	New	23	37	13	17	4	1			95
			Total		23	37	13	17	4	1			95
		Finish	New/Old?	New	11	15	2	6					34
			Total		11	15	2	6					34
	Student	Start	New/Old?	New	4	10	2	6		1	4	26	53
				Old	27	3		4				32	66
			Total		31	13	2	10		1	4	58	119
		Finish	New/Old?	New	4	4	1	1				8	18
				Old	16			3				5	24
			Total		20	4	1	4				13	42
mLG	Reading Tutor	Start	New/Old?	New	13	27	8	4	2	1			55
			Total		13	27	8	4	2	1			55
		Finish	New/Old?	New	11	22	5	3	1				42
			Total		11	22	5	3	1				42
	Student	Start	New/Old?	New	3	16	2	3	9	2	1	5	41
				Old	6	14						1	21
			Total		9	30	2	3	9	2	1	6	62
		Finish	New/Old?	New	3	12	1		3			3	22
				Old	5	9						1	15
			Total		8	21	1		3			4	37

mMH	Reading Tutor	Start	New/Old?	New	16	30	33	22	6	4			111
			Total		16	30	33	22	6	4			111
		Finish	New/Old?	New	13	16	2	1					32
			Total		13	16	2	1					32
	Student	Start	New/Old?	New	1	4		1			2	8	16
				Old	28							42	70
			Total		29	4		1			2	50	86
		Finish	New/Old?	New	1							1	2
				Old	19							12	31
			Total		20							13	33
mMW	Reading Tutor	Start	New/Old?	New	22	88	43	42	18	14			227
			Total		22	88	43	42	18	14			227
		Finish	New/Old?	New	12	19	6	4	2				43
			Total		12	19	6	4	2				43
	Student	Start	New/Old?	New	7	3				1	1	37	49
				Old	61							4	65
			Total		68	3				1	1	41	114
		Finish	New/Old?	New	2							8	10
				Old	44								44
			Total		46							8	54
mOB	Reading Tutor	Start	New/Old?	New	1	1					1		3
			Total		1	1					1		3
		Finish	New/Old?	New	1	1							2
			Total		1	1							2
	Student	Start	New/Old?	New	2								2
				Old	1								1
			Total		3								3
		Finish	New/Old?	New	1								1
			Total		1								1
mRM	Reading Tutor	Start	New/Old?	New	14	5	1	4	1		1		26
			Total		14	5	1	4	1		1		26
		Finish	New/Old?	New	10	3	1	1					15
			Total		10	3	1	1					15
	Student	Start	New/Old?	New	4								4
				Old	18								18
			Total		22								22
		Finish	New/Old?	New	4								4
				Old	13								13
			Total		17								17
mSF	Reading Tutor	Start	New/Old?	New		1	5	12	1	9			28
			Total			1	5	12	1	9			28
		Finish	New/Old?	New		1	3	7	1				12
			Total			1	3	7	1				12
	Student	Start	New/Old?	New		2	5	11	4	3		2	27
				Old				7				3	10

			Total			2	5	18	4	3		5	37
		Finish	New/Old?	New		2	4	5	1			2	14
				Old								1	1
			Total			2	4	5	1			3	15
mSK	Reading Tutor	Start	New/Old?	New	17	9	7	9	3	3			48
			Total		17	9	7	9	3	3			48
		Finish	New/Old?	New	10	3		1					14
			Total		10	3		1					14
	Student	Start	New/Old?	New	11							1	12
				Old	37								37
			Total		48							1	49
		Finish	New/Old?	New	4							1	5
				Old	16								16
			Total		20							1	21
mTB	Reading Tutor	Start	New/Old?	New	18	34	20	14	9	6	1		102
			Total		18	34	20	14	9	6	1		102
		Finish	New/Old?	New	13	13	5	3	1				35
			Total		13	13	5	3	1				35
	Student	Start	New/Old?	New	2	10	2	5	4			2	25
				Old	30	2		1				7	40
			Total		32	12	2	6	4			9	65
		Finish	New/Old?	New	1	6	2	1				1	11
				Old	16	1						5	22
			Total		17	7	2	1				6	33
mTP	Reading Tutor	Start	New/Old?	New	23	28	20	9	8	3			91
			Total		23	28	20	9	8	3			91
		Finish	New/Old?	New	13	9	6	6	1				35
			Total		13	9	6	6	1				35
	Student	Start	New/Old?	New	2	4	1					4	11
				Old	14	2						37	53
			Total		16	6	1					41	64
		Finish	New/Old?	New	2	4						1	7
				Old	7	1						22	30
			Total		9	5						23	37
mTR	Reading Tutor	Start	New/Old?	New	13	21	12	7	8		1		62
			Total		13	21	12	7	8		1		62
		Finish	New/Old?	New	12	12	1	2	1				28
			Total		12	12	1	2	1				28
	Student	Start	New/Old?	New	3	1						1	5
				Old	30							4	34
			Total		33	1						5	39
		Finish	New/Old?	New	2	1						1	4
				Old	20							2	22
			Total		22	1						3	26

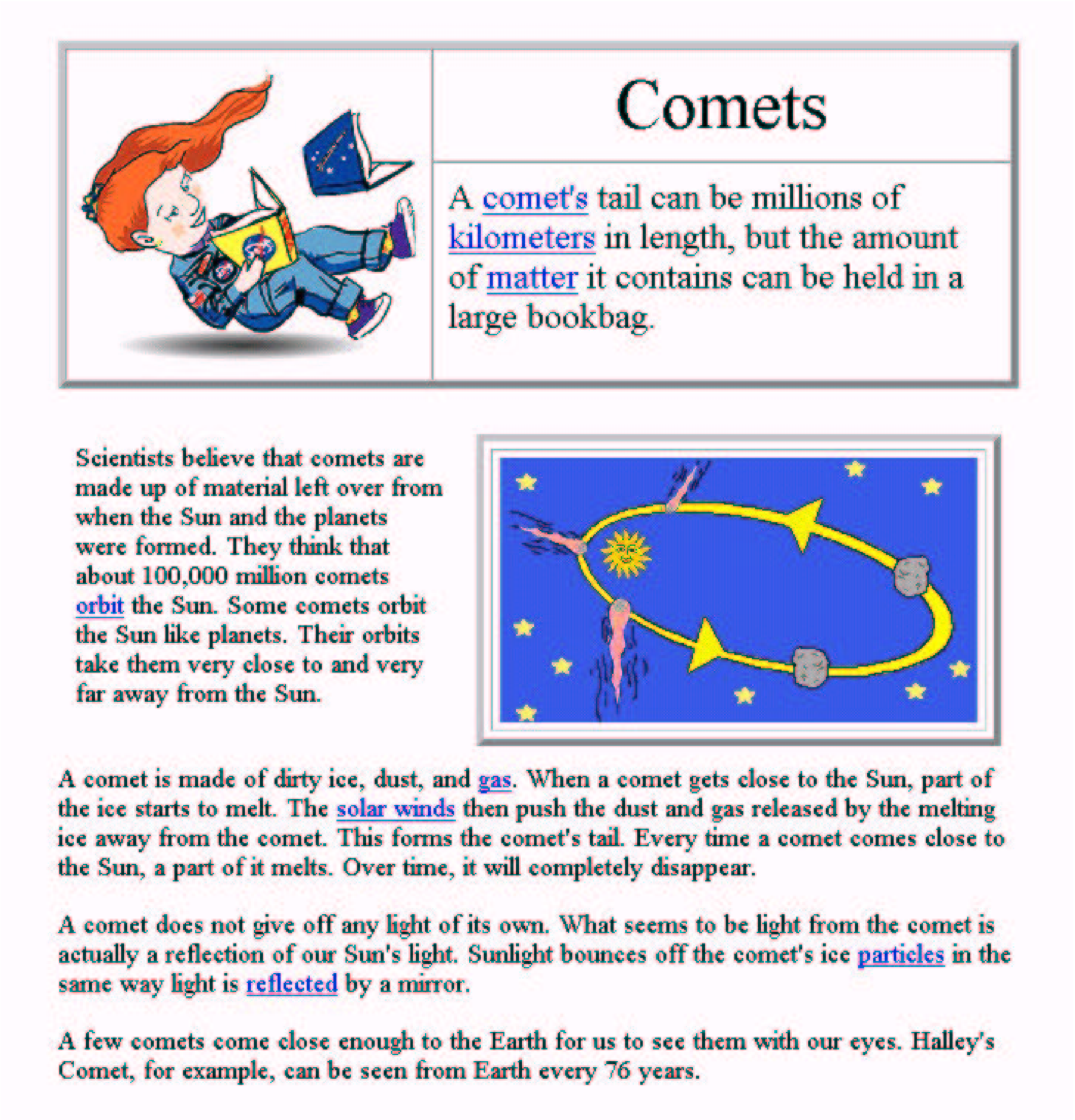
Appendix D: Comets & meteors materials, from Section 6.1

Original web pages

The next two pages contain the web pages for the comets and meteors study.

Comets

Original version downloaded in spring 2000. This version downloaded October 8, 2000 from http://starchild.gsfc.nasa.gov/docs/StarChild/shadow/solar_system_level1/comets.html.



Comets

A comet's tail can be millions of kilometers in length, but the amount of matter it contains can be held in a large bookbag.

Scientists believe that comets are made up of material left over from when the Sun and the planets were formed. They think that about 100,000 million comets orbit the Sun. Some comets orbit the Sun like planets. Their orbits take them very close to and very far away from the Sun.

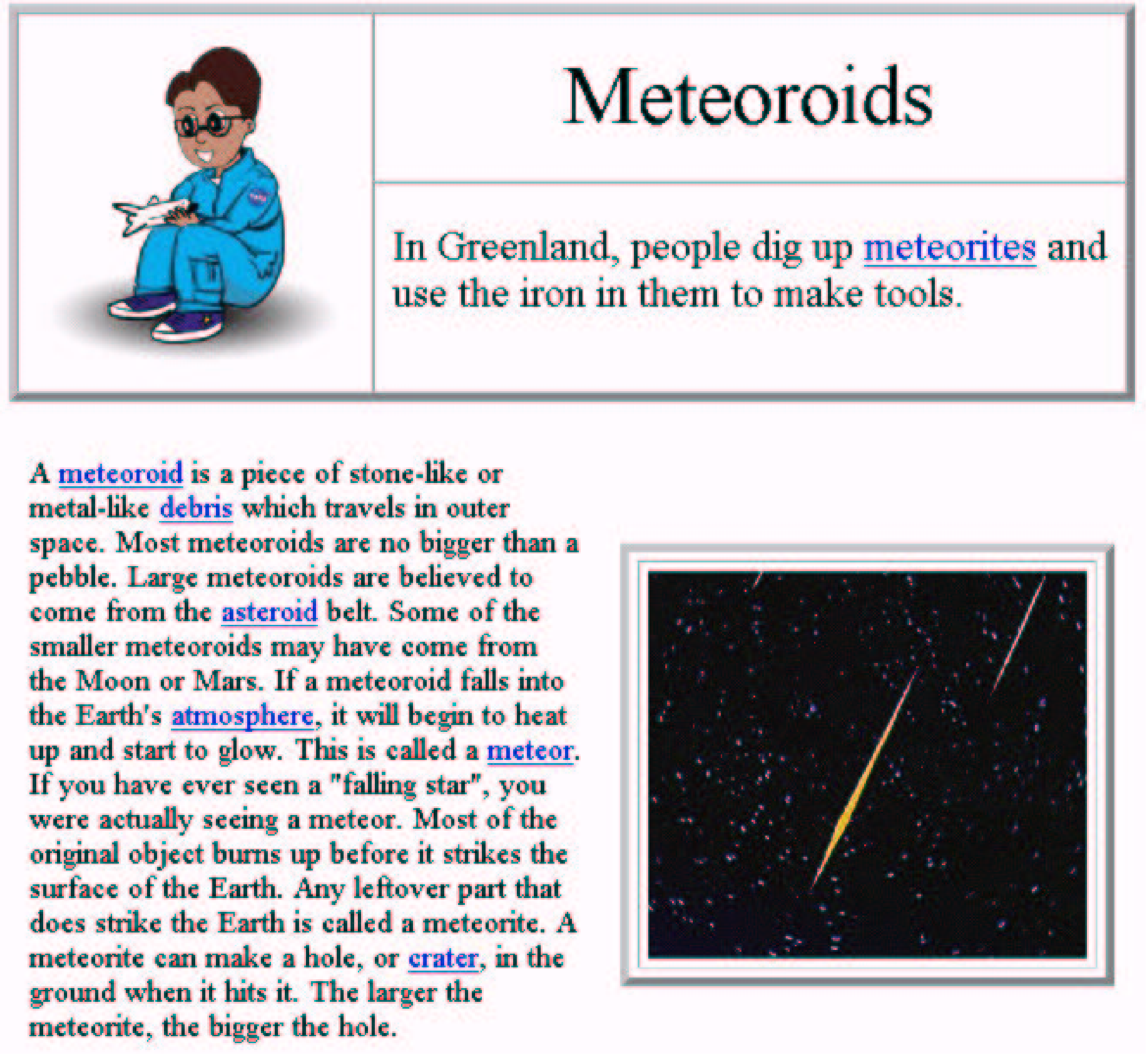
A comet is made of dirty ice, dust, and gas. When a comet gets close to the Sun, part of the ice starts to melt. The solar winds then push the dust and gas released by the melting ice away from the comet. This forms the comet's tail. Every time a comet comes close to the Sun, a part of it melts. Over time, it will completely disappear.

A comet does not give off any light of its own. What seems to be light from the comet is actually a reflection of our Sun's light. Sunlight bounces off the comet's ice particles in the same way light is reflected by a mirror.

A few comets come close enough to the Earth for us to see them with our eyes. Halley's Comet, for example, can be seen from Earth every 76 years.

Meteors

Original version downloaded in spring 2000. This version downloaded October 8, 2000 from http://starchild.gsfc.nasa.gov/docs/StarChild/solar_system_level1/meteoroids.html .



Meteoroids

In Greenland, people dig up meteorites and use the iron in them to make tools.

A meteoroid is a piece of stone-like or metal-like debris which travels in outer space. Most meteoroids are no bigger than a pebble. Large meteoroids are believed to come from the asteroid belt. Some of the smaller meteoroids may have come from the Moon or Mars. If a meteoroid falls into the Earth's atmosphere, it will begin to heat up and start to glow. This is called a meteor. If you have ever seen a "falling star", you were actually seeing a meteor. Most of the original object burns up before it strikes the surface of the Earth. Any leftover part that does strike the Earth is called a meteorite. A meteorite can make a hole, or crater, in the ground when it hits it. The larger the meteorite, the bigger the hole.

Adapted text of comets & meteors passages

Comets	Meteors
Scientists believe that comets are made up of material left over from when the Sun and the planets were formed. They think that about 100,000 million comets orbit the Sun. Some comets orbit the Sun like planets. Their orbits take them very close to and very far away from the Sun. A comet is made of dirty ice, dust, and gas. When a comet gets close to the Sun, part of the ice starts to melt. The solar winds then push the dust and gas released by the melting ice away from the comet. This forms the comet’s tail. A comet’s tail can be millions of kilometers in length, but the amount of matter it contains can be held in a large bookbag. A comet does not give off any light of its own. What seems to be light from the comet is actually a reflection of our Sun’s light. Sunlight bounces off the comet’s ice particles in the same way light is reflected by a mirror.	A meteoroid is a piece of stone-like or metal-like debris which travels in outer space. Most meteoroids are no bigger than a pebble. Large meteoroids are believed to come from the asteroid belt. Some of the smaller meteoroids may have come from the Moon or Mars. If a meteoroid falls into the Earth’s atmosphere, it will begin to heat up and start to glow. This is called a meteor. If you have ever seen a "falling star", you were actually seeing a meteor. Most of the original object burns up before it strikes the surface of the Earth. Any leftover part that does strike the Earth is called a meteorite. A meteorite can make a hole, or crater, in the ground when it hits it. The larger the meteorite, the bigger the hole. In Greenland, people dig up meteorites and use the iron in them to make tools. Sometimes, you can see more meteors than normal. That is called a meteor shower. Meteor showers take place around the same time each year.

Comets passage as augmented with explanations or nonsemantic assistance

Note: Some nonsemantic assistance contains the target word twice, to match the definition. For example, *kilometer* appears twice in the definition and thus twice in the nonsemantic control.

Text plus nonsemantic assistance	Text plus definitions
Comets COMET starts with C. Scientists believe that COMETs are made up of material left over from when the Sun and the planets were formed. ORBIT starts with O. They think that about 100,000 million comets ORBIT the Sun. Some comets orbit the Sun like planets. Their orbits take them very close to and very far away from the Sun. A comet is made of dirty ice, dust, and gas. When a comet gets close to the Sun, part of the ice starts to melt. The solar winds then push the dust and gas released by the melting ice away from the comet. This forms the comet's tail. KILOMETER starts with K. KILOMETER. A comet's tail can be millions of KILOMETERs in length, but the amount of matter it contains can be held in a large bookbag. A comet does not give off any light of its own. REFLECTION starts with R. What seems to be light from the comet is actually a REFLECTION of our Sun's light. PARTICLE starts with P. Sunlight bounces off the comet's ice PARTICLEs in the same way light is reflected by a mirror.	Comets COMET: A big ball of dirty ice and snow in outer space. Scientists believe that COMETs are made up of material left over from when the Sun and the planets were formed. ORBIT: The path followed by an object in space as it goes around another object; to travel around another object in a single path. They think that about 100,000 million comets ORBIT the Sun. Some comets orbit the Sun like planets. Their orbits take them very close to and very far away from the Sun. A comet is made of dirty ice, dust, and gas. When a comet gets close to the Sun, part of the ice starts to melt. The solar winds then push the dust and gas released by the melting ice away from the comet. This forms the comet's tail. KILOMETER: 1,000 meters. A KILOMETER equals 0.6214 miles. A comet's tail can be millions of KILOMETERs in length, but the amount of matter it contains can be held in a large bookbag. A comet does not give off any light of its own. REFLECTION: Light, heat, or sound thrown back from something. What seems to be light from the comet is actually a REFLECTION of our Sun's light. PARTICLE: A very, very tiny piece of matter such as an electron, proton, or neutron found inside of an atom. Sunlight bounces off the comet's ice PARTICLEs in the same way light is reflected by a mirror.

Meteors passage as augmented with explanations or nonsemantic assistance

Text plus nonsemantic assistance	Text plus definitions
Meteors DEBRIS starts with D. A meteoroid is a piece of stone-like or metal-like DEBRIS which travels in outer space. Most meteoroids are no bigger than a pebble. ASTEROID starts with A. ASTEROID. Large meteoroids are believed to come from the ASTEROID belt. Some of the smaller meteoroids may have come from the Moon or Mars. ATMOSPHERE starts with A. If a meteoroid falls into the Earth's ATMOSPHERE, it will begin to heat up and start to glow. METEOR starts with M. This is called a METEOR. If you have ever seen a "falling star", you were actually seeing a meteor. Most of the original object burns up before it strikes the surface of the Earth. Any leftover part that does strike the Earth is called a meteorite. CRATER starts with C. A meteorite can make a hole, or CRATER, in the ground when it hits it. The larger the meteorite, the bigger the hole. In Greenland, people dig up meteorites and use the iron in them to make tools. Sometimes, you can see more meteors than normal. That is called a meteor shower. Meteor showers take place around the same time each year.	Meteors DEBRIS: Broken, scattered remains; rubble; pieces of rubbish or litter. A meteoroid is a piece of stone-like or metal-like DEBRIS which travels in outer space. Most meteoroids are no bigger than a pebble. ASTEROID: A rocky space object that can be a few feet wide to several hundred miles wide. Most ASTEROIDs in our solar system orbit in a belt between Mars and Jupiter. Large meteoroids are believed to come from the ASTEROID belt. Some of the smaller meteoroids may have come from the Moon or Mars. ATMOSPHERE: All the gases which surround a star, like our Sun, or a planet, like our Earth. If a meteoroid falls into the Earth's ATMOSPHERE, it will begin to heat up and start to glow. METEOR: An object from space that becomes glowing hot when it passes into Earth's atmosphere. This is called a METEOR. If you have ever seen a "falling star", you were actually seeing a meteor. Most of the original object burns up before it strikes the surface of the Earth. Any leftover part that does strike the Earth is called a meteorite. CRATER: A hole caused by an object hitting the surface of a planet or moon. A meteorite can make a hole, or CRATER, in the ground when it hits it. The larger the meteorite, the bigger the hole. In Greenland, people dig up meteorites and use the iron in them to make tools. Sometimes, you can see more meteors than normal. That is called a meteor shower. Meteor showers take place around the same time each year.

Matching task for comets story

Please match each word with its definition
by drawing a line as shown below:

reflection	A small piece of something
kilometer	Frozen water
particle	A mirror image
comet	A path around the Sun
ice	A ball of ice in space with a tail
orbit	About half a mile

Matching task for meteors story

Please match each word with its definition
by drawing a line as shown below:

iron	A rock in space
asteroid	A rock falling from space to the Earth
debris	A hard metal
atmosphere	A hole made by a rock
crater	Junk or pieces of rock
meteor	The air around the Earth

Appendix E: Materials from limericks study, in Section 6.2

Limericks containing target words

There was an Old Man of Cape Horn,
Who wished he had never been born;
So he sat on a chair,
Till he died of despair,
That dolorous Man of Cape Horn.

There was an old person of Wick,
Who said, ‘Tick-a-Tick, Tick-a-Tick;
Chickabee, Chickabaw,’
And he said nothing more,
That laconic old person of Wick.

There was an Old Person of Chili,
Whose conduct was painful and silly;
He sate on the stairs,
Eating apples and pears,
That imprudent Old Person of Chili.

There was an old man of Hong Kong,
Who never did anything wrong;
He lay on his back,
With his head in a sack,
That innocuous old man of Hong Kong.

There was an Old Person of Gretna,
Who rushed down the crater of Etna;
When they said, “Is it hot?”
He replied, “No, it’s not!”
That mendacious Old Person of Gretna.

There was an Old Lady of Prague,
Whose language was horribly vague;
When they said, “Are these caps?”
She answered, “Perhaps!”
That oracular Lady of Prague.

There was an Old Person of Bangor,
Whose face was distorted with anger;
He tore off his boots,
And subsisted on roots,

That irascible person of Bangor.

There was an old person of Loo,
Who said, 'What on earth shall I do?'
When they said, 'Go away!' –
She continued to stay,
That vexatious old person of Loo.

Definitions for target words

We can say someone is dolorous if they are mournful, or feel really bad.

We can say someone is laconic if they say something brief or to the point.

We can say someone is imprudent if they are unwise, or do things they shouldn't do.

We can say someone is innocuous if they don't hurt anybody and don't put anyone in danger.

We can say someone is mendacious if they don't tell the truth or say something false.

We can say someone is oracular if they tell you things that are puzzling or hard to understand.

We can say someone is irascible if they easily get upset and angry.

We can say someone is vexatious if they keep bothering you and won't leave you alone.

Tests for limericks study

The tests for the limerick study consisted of two questions on each of eight words. The next two pages contain the tests.

Your Name: ____________________

Here are a few questions about some hard words.
It's okay if you don't know all the words. Just do your best.
Please draw a circle around your answer, like this:

Have you ever seen the word ***giant*** before? [✓ Yes] No

If someone is ***giant*** they must be…

cold [✓ big] nice fast

Thanks!

----

1. Have you ever seen the word ***dolorous*** before? Yes No

If someone is ***dolorous*** they must be…

angry sad tired afraid

2. Have you ever seen the word ***laconic*** before? Yes No

If someone is ***laconic*** they say things that are…

short loud wrong boring

3. Have you ever seen the word ***imprudent*** before? Yes No

If someone is ***imprudent*** they must be…

slow quiet tall foolish

4. Have you ever seen the word ***innocuous*** before? Yes No

If someone is ***innocuous*** they must be…

worried quick harmless ready

5. Have you ever seen the word ***mendacious*** before? Yes No

If someone is ***mendacious*** they must be…

smart careful friendly lying

6. Have you ever seen the word ***oracular*** before? Yes No

If someone is ***oracular*** they must be…

unclear mean super happy

7. Have you ever seen the word ***irascible*** before? Yes No

If someone is ***irascible*** they must easily get…

curious tired mad silly

8. Have you ever seen the word ***vexacious*** before? Yes No

If someone is ***vexacious*** they must be…

friendly annoying lucky pretty