Evaluation of an Automated Reading Tutor that Listens: Comparison to Human Tutoring and Classroom Instruction

Jack Mostow, Greg Aist, Paul Burkhead, Albert Corbett, Andrew Cuneo,
Susan Eitelman, Cathy Huang, Brian Junker, Mary Beth Sklar, and Brian Tobin
Project LISTEN, 4213 NSH, Carnegie Mellon University, Pittsburgh, PA 15213
(412) 268-1330 voice / 268-6436 FAX
http://www.cs.cmu.edu/~listen
Mostow@cs.cmu.edu

Revised 15 August 2002. To appear in *Journal of Educational Computing Research, 29*(1).

Abstract

A year-long study of 131 second and third graders in 12 classrooms compared three daily 20-minute treatments. (a) 58 students in 6 classrooms used the 1999-2000 version of Project LISTEN's Reading Tutor, a computer program that uses automated speech recognition to listen to a child read aloud, and gives spoken and graphical assistance. Students took daily turns using one shared Reading Tutor in their classroom while the rest of their class received regular instruction. (b) 34 students in the other 6 classrooms were pulled out daily for one-on-one tutoring by certified teachers. To control for materials, the human tutors used the same set of stories as the Reading Tutor. (c) 39 students served as in-classroom controls, receiving regular instruction without tutoring. We compared students' pre- to post-test gains on the Word Identification, Word Attack, Word Comprehension, and Passage Comprehension subtests of the Woodcock Reading Mastery Test, and in oral reading fluency.

Surprisingly, the human-tutored group significantly outgained the Reading Tutor group only in Word Attack (main effects $p<.02$, effect size .55). Third graders in both the computer- and human-tutored conditions outgained the control group significantly in Word Comprehension ($p<.02$, respective effect sizes .56 and .72) and suggestively in Passage Comprehension ($p=.14$, respective effect sizes .48 and .34). No differences between groups on gains in Word Identification or fluency were significant. These results are consistent with an earlier study in which students who used the 1998 version of the Reading Tutor outgained their matched classmates in Passage Comprehension ($p=.11$, effect size .60), but not in Word Attack, Word Identification, or fluency.

To shed light on outcome differences between tutoring conditions and between individual human tutors, we compared process variables. Analysis of logs from all 6,080 human and computer tutoring sessions showed that human tutors included less rereading and more frequent writing than the Reading Tutor. Micro-analysis of 40 videotaped sessions showed that students who used the Reading Tutor spent considerable time waiting for it to respond, requested help more frequently, and picked easier stories when it was their turn. Human tutors corrected more errors, focussed more on individual letters, and provided assistance more interactively, for example getting students to sound out words rather than sounding out words themselves as the Reading Tutor did.

Introduction

"Research also is needed on the value of speech recognition as a technology … in reading instruction." (NRP, 2000)

Literacy is more important than ever in today's high-tech economy. Unfortunately, the National Assessment of Educational Progress (NCES, 2000) shows that a distressingly high percentage of the nation's children read less proficiently than they should – a picture that has shown little change in 30 years. For example, the 2000 Nation's Report Card showed that 37% of fourth graders read below the Basic level, and only 32% read at or above the Proficient level. Although "higher-performing students have made gains" since 1992, "the score at the 10th percentile was lower in 2000 than it was in 1992. This indicates that lower-performing students have lost ground" (http://nces.ed.gov/nationsreportcard/reading/results/scalepercent.asp). To raise literacy to the levels

required for the 21st century, K-3 education must become radically more productive than one-to-many classroom instruction in the tradition of the 19th century.

Studies of one-on-one literacy tutoring have demonstrated dramatic improvements, as summarized by the Committee for Preventing Reading Difficulties in Young Children in (Snow, Burns, & Griffin, 1998). Some key lessons can be drawn from this research:

1. Effective individual tutoring involves spending extra time on reading – typically 30 minutes daily for much or all of a school year. Thus individual tutoring is an expensive proposition.
2. Although extra time may be necessary, it is not sufficient; not all tutoring programs are effective, especially for certain kinds of reading difficulties.
3. Tutor effectiveness depends on training and supervision of tutors – another considerable expense.
4. Student response to tutoring needs to be monitored closely by assessing student progress.
5. A key element of effective tutoring is reading connected, engaging text. Extensive assisted oral reading of connected text has been shown to improve overall reading ability (Cunningham & Stanovich, 1991; Fielding, Wilson, & Anderson, 1986; Leinhardt, Zigmond, & Cooley, 1981; Nagy, Herman, & Anderson, 1985) – not just word identification, but more general cognitive processing and accumulation of background knowledge (Cunningham & Stanovich, 1991).
6. Other activities common to effective tutoring include word study and writing. However, the cause-and-effect connections between tutorial activities and student gains are not clearly understood.
7. Gains by tutored children compared to control groups may persist on measures specific to the treatment, yet without extending to other aspects of reading performance.

In short, individual human tutoring is expensive, and often – but not always – provides lasting benefits. Fortunately, the same advances in technology that make literacy gains imperative may also provide a powerful and cost-effective tool to help achieve them – namely, automated individual literacy tutoring (Snow, Burns, & Griffin, 1998). But current literacy software is not always effective. Moreover, commercially available educational software lacks a key element of effective human tutoring: it doesn't listen to the student read connected text. This limitation prevents the software from detecting oral reading difficulties on the part of the reader. Instead, the software assumes that readers will ask for help when they need it. However, studies of spoken assistance on demand (Lundberg & Olofsson, 1993; McConkie, 1990; Olson, Foltz, & Wise, 1986; Olson & Wise, 1987) have revealed a serious flaw in assuming that young readers are willing and able to ask for help when they need it. Children with reading difficulties often fail to realize when they misidentify a word. This problem is especially acute for children with weak metacognitive skills.

Previous work on Project LISTEN: To address the limitations of previous reading software, Project LISTEN has developed (and continues to improve) an automated Reading Tutor that listens to children read aloud, helps them, and also lets them write and narrate stories. The Reading Tutor uses speech recognition to analyze children's disfluent oral reading (Aist & Mostow, 1997a; Mostow & Aist, 1999a, c; Mostow, Hauptmann et al., 1993; Mostow, Roth et al., 1994). Its design is modelled after expert reading teachers, based on research literature, and adapted to fit technological capabilities and limitations (Mostow & Aist, 1999b; Mostow, Roth et al., 1994). Along the way we have evaluated successive prototypes (Aist & Mostow, 2000, in press; Aist, Mostow et al., 2001; Mostow & Aist, 2001; Mostow, Aist et al., in press; Mostow, Roth et al., 1994). For details of these different aspects, please see the cited publications; we now summarize this prior work just enough to place the current study in context.

Project LISTEN's initial studies observed expert tutoring and used "Wizard of Oz" experiments to simulate automated assistance modelled after it. These experiments supported the iterative design of the "look and feel" for such assistance. A within-subject study of 12 low-reading second graders (Mostow, Roth *et al.*, 1994) showed that this assistance enabled them to read and comprehend material at a level 6 months higher than they could read on their own.

Replacing the human "wizard" in these experiments with interventions triggered by a speech recognizer yielded an automated "Reading Coach." A May 1994 within-subject study of 34 second graders (Mostow & Aist, 2001) showed that they averaged 20% higher comprehension on a third-grade passage with the Reading Coach's automated assistance than without. Both these studies measured assistive effects, not gains. That is, they just compared how well students read with help versus without help. In contrast, our subsequent experiments tested whether such assistance helped students learn over time. Redesigning, scaling up, and "kid-testing" the Reading Coach to support extended use on school-affordable personal computers yielded a new program called the Reading Tutor.

A 1996-97 pilot study (Aist & Mostow, 1997b) at a public elementary school in a low-income, predominantly African-American inner-city community in Pittsburgh, Pennsylvania, included six of the lowest third graders, who started almost three years below grade level. Using the Reading Tutor under individual supervision by a school aide, they averaged two years' progress in less than eight months, according to informal reading inventories administered by school personnel in October 1996 and June 1997. The school was excited by these results because even 8 months' progress in 8 months' time would have been a dramatic improvement for these students.

To enable children to operate the Reading Tutor independently under regular classroom conditions, we added child-friendly mechanisms for logging in and picking which stories to read. We also expanded the Reading Tutor's repertoire of spoken and graphical interventions (Mostow & Aist, 1999b) to sound out, syllabify, rhyme, spell, hint, prompt, preempt likely mistakes, interrupt (Aist, 1998), encourage (Aist & Mostow, 1999), and praise.

In spring 1998, a four-month within-classroom controlled study at the same school compared the Reading Tutor, regular instruction, and commercial reading software. We summarize this study here and in Table 10; for details, see (Mostow, Aist *et al.*, in press). All 72 students in 3 classrooms (grades 2, 4, and 5) that had not previously used the Reading Tutor were independently pre-tested on the Word Attack, Word Identification, and Passage Comprehension subtests of the Woodcock Reading Mastery Test (Woodcock, 1987), and on oral reading fluency. We split each class into 3 matched treatment groups – Reading Tutor, commercial reading software, or regular classroom activities, including other software use. We assigned students to treatments randomly, matched within classroom by pretest scores. All treatments occurred in the classroom, with one computer for each treatment.

Table 1 shows the results based on the WRMT-NU/Revised norms. (The analysis in (Mostow, Aist *et al.*, in press) is based on older norms but yielded qualitatively similar results.) Even though the study lasted only 4 months, and actual usage was a fraction of the planned daily 20-25 minutes, the 22 students who used the 1998 version of the Reading Tutor gained more in Passage Comprehension than their 20 classmates in the control group, and progressed faster than their national cohort. No other between-treatment differences in gains were significant. The difference in comprehension gains was suggestive at $p = 0.106$ using ANCOVA with pretest score as a covariate, effect size 0.60. For the 17 matched pairs, the difference was significant at $p < 0.002$ on a 2-tailed paired T-test, with effect size 1.52. As the principal said, "these children were closing the gap."

<<insert near here: Table 1: Results of spring 1998 4-month within-classroom comparison>>

The 1998 study suggested several lessons. First, the Reading Tutor seemed to help younger grades and weaker students more, but the sample was too small to make these interactions statistically significant. Second, although the within-classroom design controlled for teacher effects, it let one treatment affect another. In particular, equity concerns led teachers to equalize computer time among all three treatment groups, thereby giving students in the "regular classroom activities" treatment more computer time than they might otherwise have gotten. Third, we noticed that poor readers tended to pick the same easy stories over and over. To address this behavior, we subsequently redesigned the Reading Tutor to take turns with the student at picking stories. Analysis of recorded story choices in successive versions of the Reading Tutor confirmed that story choice was now measurably more efficient and effective (Aist & Mostow, 2000, in press).

Experimental Design

In 1999-2000, to "prove and improve" the Reading Tutor – that is, to evaluate against conventional instruction, and to identify areas for improvement – we compared it to one-on-one human tutoring, and to spending the same time in regular classroom activity. We now describe the study design in terms of the students who participated, the treatments they received, and the outcome measures we used. We describe the three treatments in terms of *setting*, *personnel*, *materials*, and *activities*. Later we analyze outcome and process variables, and summarize finer-grained evaluations performed as part of the study and published in more detail elsewhere.

Students: To exclude effects of prior Reading Tutor use, we recruited an urban elementary school in a small city near Pittsburgh, Pennsylvania, that had not previously used the Reading Tutor. Its student body was mixed-income, with 75% qualifying for free or reduced school lunch. Approximately 65% were white, and 35% African-American. Based on the 1998 study, which suggested that the Reading Tutor made a bigger difference for poorer and younger readers, we decided to focus on bottom-half students in grades 2 and 3. To reduce the amount of individual pretesting required, we asked teachers in 12 second and third grade classrooms to each choose their 12 poorest readers, rather than pretest their entire classes to decide which students to include. The resulting study sample of 144 second and third graders (ranging from 7 to 10 years old) focussed on the population that we believed the Reading Tutor had the greatest potential to help, and offered greater statistical power than the 1998 study. 131 of the 144 students completed the study.

Assignment to treatment: We initially assigned 60 students to use the Reading Tutor, 36 students to human tutors, and 48 students to the control condition. We assigned each student to the same treatment 20 minutes daily for the entire school year, so as to maximize the power of the study to resolve differences between treatments. Each classroom had only one type of tutoring, so as to keep either type from influencing the other. For example, we didn't want children who used the Reading Tutor to see paper copies of the same stories, lest it distort the Reading Tutor's assessment of their oral reading fluency.

Six students was the most that each human tutor could cover, given her other duties. So each tutor tutored 6 students from one class, one at a time, with occasional substitutions due to other responsibilities. The other 6 students in the same room served as in-room controls, chosen by stratified random selection so as to make treatment groups statistically well-matched.

We wanted to maximize the number of students who used the Reading Tutor, in hopes of learning more about what kinds of students it helped most. Ten students was the maximum we thought could share one Reading Tutor. The reason is that ten 20-minute sessions add up to just over 3 hours, which is about the maximum Reading Tutor usage feasible in a classroom during the school day. The rest of the time the class is out of the room or engaged in special subjects. Accordingly, we assigned 10 of the 12 study subjects in each classroom to use the Reading Tutor, and the other 2 as in-room controls, randomly choosing one from the top 6 and one from the bottom 6, based on their Total Reading Composite pretest scores on the Woodcock Reading Mastery Test (Woodcock, 1998).

The resulting assignment of students can be summarized as follows. In the 6 rooms with a human tutor, 6 students were tutored, and 6 were controls. In the 6 rooms with a Reading Tutor, 10 students used it, and 2 were controls. Two teachers tried to put one or both of the in-room controls on the Reading Tutor, but could not always get them on. We excluded these three "part-timers" from analysis.

Setting: Regular instruction took place in classrooms, with class size of about 24. Individual human tutoring took place at a desk in the hall outside the classroom. As in the 1998 study, students took turns throughout the school day using one Reading Tutor computer in their classroom. This implementation avoided the cost of staffing a separate lab, but required considerable teacher cooperation.

Personnel: Treatment "personnel" included classroom teachers, human tutors, and the Reading Tutor. According to the principal, all classroom teachers in the study were comparably experienced veteran teachers. Teacher cooperation was essential to classroom use of the Reading Tutor, so the principal chose six classrooms to get Reading Tutors based on his estimate of teachers' willingness to cooperate – possibly a confound, but necessary. The human tutors were certified elementary teachers already employed by the school. Studies of one-on-one tutoring in 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 (Clay, 1991). Using certified teachers rather than paraprofessionals has been associated with positive results for one-on-one reading tutoring (Wasik & Slavin, 1993). The tutors in our study had at least a bachelor's degree in elementary education and 0-2 years experience teaching (often preschool children), but no specialized training in reading tutoring. Thus we expected them to do better than classroom instruction, but not as well as the world's best tutor – which would have been an unrealistic comparison even for a research study, let alone for large-scale implementation. The Reading Tutor was the version of 9/1/99, with changes confined to a single mid-year patch that addressed Y2K issues and fixed a few bugs without altering user-visible functionality.

Materials: The text used in reading instruction and practice is an important variable. Regular instruction used a basal reading curriculum. To control for materials across the two tutoring conditions, we asked human tutors to use the same set of stories as the Reading Tutor, to refrain from bringing in outside books, and to limit any writing (by student or tutor) to student journals we designed for that purpose. We gave the tutors bound copies of the Reading Tutor stories at the start of the year. After using the Reading Tutor for a few months some students started running out of new material to read, so in February 2000 we added more stories to the Reading Tutor and gave them to the human tutors in the form of a supplemental volume.

The stories came from various sources, including *Weekly Reader*, Project Gutenberg (www.gutenberg.net/) and other public-domain Web sources, and stories authored in-house. Each story had a (human-assigned) level. Levels K, A, B, C, D, and E corresponded to kindergarten through grade 5. Each level had about two dozen stories, ranging in length from a few dozen words in level K stories to several hundred words in level E stories. Level K had stories like "Bob got a dog" with a few, short, decodable sentences. Level A had letter stories like "The letter A" ("APPLE starts with A…."), letter-sound stories like "The first sound in CHIN" and "The vowel sound in MOON," nursery rhymes like "Jack and Jill," and some *Weekly Reader* stories. Level B had Aesop's fables like "The Wolf in Sheep's Clothing," arithmetic tables like "Dividing By 3," poems like "Eletelephony, by Laura Richards," a few more advanced letter-sound stories like "The vowel sound in LAKE, BRAID and TABLE," and *Weekly Reader* stories. Level C had poems, fables, *Weekly Reader* stories, an excerpt of Martin Luther King's "I have a dream" speech, and stories like "Why do dogs chase cats?" from the National Science Foundation's "Ask a Scientist or Engineer" website (www.nsf.gov/nstw_questions/). Level D consisted mostly of longer stories split into installments like "Beauty And The Beast, Part 3" and "The Adventures of Reddy Fox, part 4." Level E consisted mostly of installments from "Dorothy and the Wizard in Oz" and "Alice in Wonderland." In addition, the Reading Tutor had level H for stories on how to use the Reading Tutor, and level U for students to author their own stories. The printed stories omitted levels H and U.

Activities: Instruction and practice may use the same text materials in very different ways. The extent to which we were able to record and characterize them differed by treatment. Classroom reading instruction typically involves a wide range of whole-class, small-group, and individual activities that varies by teacher, and hence can be characterized only imperfectly, even with extensive classroom observation beyond the scope of this study. However, we did use a questionnaire to ask each teacher how much time she spent each day on scheduled reading instruction and on additional reading instruction and practice, such as when reading science and social studies materials. Teachers reported from 50 to 80 minutes of scheduled reading instruction per day. The amount of time spent on additional reading-related activities varied more widely across classes, depending on the teacher's definition of "reading-related activities," from 20 minutes per day up to 270 minutes per day for one teacher who reported that "children are always engaged in some aspect of the reading process." In spite – or even because – of its variability, "current practice" has face validity as a baseline against which to compare any proposed treatment. Moreover, its ill-defined, idiosyncratic nature is somewhat obviated by the fact that students in all three treatment

groups received mostly the same instruction. Students in both tutoring conditions averaged 0 to 15 minutes more time per day on assisted reading and writing, depending on how much of their tutoring occurred during language arts time. Teachers rotated scheduling to vary which subjects students missed while being tutored.

Tutors helped students read and write. Human tutors vary, just as teachers do. Thanks to using a prespecified set of stories and restricting all writing to student journals, tutors were able to log the activities performed in each day's sessions on a 1-page form with the date and tutor's name at the top, and a few lines for each student session. The tutor identified each story by its level, page number, and brief title, and recorded which pages the student read, and whether the story was too easy, OK, or too hard. Writing activities were listed as level W, with page numbers referring to the bound writing journal we provided for each student. Entering the tutor logs into database form yielded a comprehensive, machine-analyzable summary of the human tutors' activities in the entire year's sessions – all 2,247 of them, with a total of 6,427 activities. The student journals provided a similarly comprehensive record of the writing activities, albeit in a form requiring human interpretation to decipher, and that did not record the tutorial interventions that produced the writing. We used a digital camera to capture the session logs and student journals on site during the study. We also videotaped some tutoring sessions in order to complement the comprehensive but coarse-grained logs and the more detailed but written-only journals. We coded the videotapes for several process variables (described later) to characterize aspects of tutor-student interactions not captured by the comprehensive data.

The (1999-2000 version of the) Reading Tutor provided computer-assisted oral reading of connected text, as described in more detail elsewhere (Mostow & Aist, 1999b). Each session consisted of logging in, answering multiple choice questions about any vocabulary words introduced in the previous session (Aist, 2001b), and then reading or writing stories. To keep poor readers from rereading the same easy stories over and over, the Reading Tutor took turns with the student at picking which story to read next (Aist & Mostow, 2000, in press). The Reading Tutor chose previously unread stories at the student's estimated reading level, and invited the student to pick stories at that level too. When it was their turn, students could pick a story at any level to read or reread, or choose to type in and narrate a story (Mostow & Aist, 1999a, c) that other children could then pick (and sometimes did). The Reading Tutor deliberately under-estimated a student's initial reading level based on age, to avoid frustrating children with stories at too high a level. It then adjusted its estimate up or down if the student's assisted reading rate on a previously unread story fell above 30 wpm or below 10wpm, as described in (Aist, 2000; Aist & Mostow, in press). The Reading Tutor displayed the chosen story in a large font, adding one sentence at a time. The Reading Tutor listened to the child read the sentence aloud, going on to display the next sentence if it accepted the reading or the child clicked an on-screen Go button. The Reading Tutor intervened if it detected a serious mistake, a skipped word, a long silence, a click for help, or a difficult word. It also gave occasional praise for good or improved performance.

The Reading Tutor chose from a repertoire of interventions at different grain sizes. To appear animate and attentive, it displayed a persona that blinked sporadically and gazed at the cursor position or whichever word it expected to hear next, which it also highlighted with a moving shadow. Both gaze and shadow responded visibly to oral reading. To encourage the student to continue reading, it occasionally made a backchannelling sound like "uh-huh" when the student hesitated for two seconds. To call attention to a skipped word, it underlined the word, sometimes with a slight coughing sound. To give help on a word, the Reading Tutor selected from among several forms of assistance. It could speak the word aloud; recue the word by reading the words that led up to it; decompose the word into syllables, onset and rime, or phonemes; compare it to a word with the same onset or the same rime; or (rarely) display a picture or play a sound effect. In general, when more than one intervention was possible and reasonable, the Reading Tutor chose one of them at random, so as to provide variety both for the sake of interest and to generate data on their relative efficacy. To explain a new word, the Reading Tutor sometimes presented a short, automatically generated "factoid" about the word for the student to read (with assistance) just before the sentence containing the word, as reported in more detail elsewhere (Aist, 2001a, b, 2002a). To read a sentence aloud, the Reading Tutor randomly played back either a fluent narration of the entire sentence or else a recording, of, each, successive, word, one, at, a, time, like, this. It provided such whole-sentence help when the student requested it by clicking, when the student missed more than one content word in the sentence, when the student read with long hesitations, or sometimes pre-emptively when the sentence contained hard words. To

prompt a student who got stuck, the Reading Tutor said to read aloud or click for help, or it read the sentence itself. To praise the student without knowing for sure which words were read correctly, the Reading Tutor played a recording of a child or adult saying something encouraging but unspecific, such as "you're a great reader!" For intelligibility, expressiveness, and personality, the Reading Tutor used digitized human speech recorded by various adults and children, resorting to synthesized speech only for occasional words of which it had no recording.

One advantage of technology is its super-human ability to collect copious data. The Reading Tutor recorded data in several forms, which we now enumerate in increasing order of detail. The class roster displayed on the Reading Tutor between sessions was intended to help teachers and students monitor scheduled usage and student performance. The roster was modelled in part after charts that teachers had made to facilitate scheduling. It showed how long each student had read that day, with a blank next to students who had not yet read that day, e.g.:

17 min.	Danielle Thomas	New stories: 38	New words: 700
	Timesha Peterson	New stories: 29	New words: 479

The roster displayed the count of distinct stories and words each student had seen to date. Based on previous experience we expected students to compete on anything that looked like a score, so we displayed numbers that would encourage students to read challenging new stories rather than try to rack up "points" by rereading old stories. Clicking on the student's story count in the roster brought up the student portfolio, which listed each story the student started reading, on what date, who chose the story (student or Reading Tutor), the story level, whether s/he finished reading the story, how many times the student had finished that story before, and the title of the story. Clicking on the student's word count brought up the student's word list, which listed individual words the student encountered, on what date, who chose the story (student or Reading Tutor), the story level, the number of times the student had finished that story before, and the title of the story. Every student utterance was digitally recorded in a separate file, with parallel files showing the sentence the student was supposed to read, and the time-aligned sequence of words output by the speech recognizer. A database recorded events the Reading Tutor needed to remember, such as finishing a story for the nth time, or encountering a new word. An excruciatingly detailed log recorded, millisecond by millisecond, the timestamped sequence of internal events in the Reading Tutor, for later debugging and analysis. We used these various data sources and the human tutors' logs to compare the same process variables for different tutors, as we shall soon describe. But before we compare tutoring processes, we first evaluate how well they worked.

Outcomes

We now compare the three study treatments – baseline instruction, the Reading Tutor, and human tutors. To gather results comparable to other published studies on reading instruction, we used the Woodcock Reading Mastery Test (WRMT) (Woodcock, 1998), an individually administered reading test. The WRMT consists of several subtests, each of which tests a specific area of reading skill. A pre- to post-test gain in *raw* score indicates progress in absolute terms. Each WRMT subtest is normed relative to a national sample to have a mean of 100 and a standard deviation of 15. The WRMT norms scores not only by grade, but by month within grade. Thus a gain of 0 in normed score means that the student stayed at the same percentile relative to his or her peers.

Trained testers pre-tested students in September 1999 and post-tested them in May 2000. This study used four WRMT subtests. Word Attack (WA) measures decoding skills by testing the ability to decipher rare or non-words. Word Identification (WI) tests the ability to read individual words out loud. Word Comprehension (WC) tests if the student understands individual words well enough to supply an antonym or synonym, or to complete an analogy. Passage Comprehension (PC) tests the ability to understand a 1-2 sentence cloze passage well enough to fill in the missing word. Total Reading Composite (TRC) combines these four subtests into a single overall score. We used one measure in addition to the WRMT. Fluency (FLU) measures independent oral reading fluency as the median number of words read correctly in one minute for each of three prespecified passages. Fluency offers the advantages of curriculum-based measurement and correlates highly with comprehension (Deno, 1985). This unassisted oral reading rate was measured both on passages at the student's grade level and (where different) on passages at the student's reading level.

We now compare pre- to post-test gains by the three treatment groups on the four WRMT subtests we used (Word Attack, Word Identification, Word Comprehension, and Passage Comprehension) and oral reading fluency. We address the following questions in turn: Which treatment groups improved? Which treatment groups improved faster than their national cohort? Did treatment group outcomes differ? Did tutoring help? Did individual tutors differ?

Which treatment groups improved? Raw scores rose significantly from pre- to post-test for all three treatment groups in both grades on every WRMT subtest and on fluency. To check significance, we used a T-test to compare individual raw gains (post-test minus pretest) against the constant zero. This comparison is equivalent to a repeated measures comparison of pre- to post-test scores. All improvements were significant at p=.000 except for grade 3 Word Attack, which was significant at p=.026 for the control group, p=.007 for the Reading Tutor group, and p=.001 for human tutoring. However, gains in raw scores are not surprising because they reflect children's general growth over the year. To filter out such general growth, we next compared to national norms.

Which treatment groups gained more than their national cohort? To answer this question, we looked at normed gains on the four WRMT subtests (omitting fluency because the fluency test was not normed). 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. We used 2-tailed T-tests to compare gains on the four WRMT subtests to zero. A gain significantly greater than zero represents progress significantly faster than the norming sample.

With one exception, in grade 2 all three treatment groups significantly outgained the norms in Word Attack ($p<.03$) and Word Comprehension ($p<.04$) but not in Word Identification ($p>.35$) or Passage Comprehension ($p>.15$). The exception is that the Reading Tutor's 3-point normed gain in Word Attack was not significant ($p=.2$).

In grade 3, the control group did not significantly outgain the norms on any of the WRMT subtests ($p>.45$). The human tutor group significantly outgained the norms on Word Identification ($p=.003$), Word Comprehension ($p=.01$), and Passage Comprehension ($p<.02$), though not Word Attack ($p>.2$). The Reading Tutor group outgained the norms significantly on Word Comprehension ($p=.001$) and Passage Comprehension ($p=.009$) and marginally on Word Identification ($p<.08$), but on Word Attack gained marginally *less* (by 3 points) than the norms ($p<.07$).

Did treatment group outcomes differ? We wanted to identify significant differences among treatment groups on each outcome measure. We used analysis of variance of gains by treatment and grade, with an interaction term for grade and treatment, and pretest scores as covariates. Standard exploratory data analysis methods identified a few significant and influential outliers in gain scores. Since we are interested in the typical effect of independent variables on gains, it is important to control for these gross outliers. Rather than deplete sample sizes by removing these data points, we Winsorized our sample at the 1st and 99th percentiles. To compare gains between similarly skilled students, we had randomized the assignment of students to treatments, stratified by Total Reading Composite pretest score. To further control for students' pretest differences on individual subtests, our models included pretest scores as covariates. But *which* pretest scores? To maximize the fit of the model for each outcome gain to the data, we searched through the set of combinations of possible covariates (Word Attack, Word Identification, Word Comprehension, Passage Comprehension, and fluency) and minimized the error remaining between the model and the actual data. Correlations between pretest scores and gains were generally negative, indicating regression to the mean and/or successful remediation of student deficits. However, regression to the mean cannot explain *differences* in gains between treatment groups. Where we found significant effects of treatment, we computed the effect size of the difference between two treatment groups as the difference between the means of their adjusted gains, divided by their average standard deviation. Where grade interacted significantly with treatment, we analyzed grade 2 and grade 3 separately. However, for clarity we report all results by grade. Table 2 summarizes the results, including significance levels for main effects.

<<insert near here: Table 2: Results of 1999-2000 8-month comparison of treatment groups' pretest scores and gains on each test, by grade>>

As Table 2 shows, we found surprisingly few significant differences among treatments. We expected the human tutors to lead across the board. Instead, human tutoring significantly outgained the Reading Tutor only in Word Attack (p=.02, ES=.55). Human and computer tutoring both surpassed the control in grade 3 Word Comprehension gains (p=.02, ES = .56 and .72, respectively). In grade 3 Passage Comprehension, a trend favored the Reading Tutor over the control (p=.14, ES=.48). No other differences were significant. The absence of significant differences in fluency gains is especially surprising, because fluency is considered a sensitive measure of growth (Fuchs, Fuchs et al., 1993).

Did tutoring help? Treatment condition in this study was partly correlated with classroom, so treatment group effects may depend both on the treatment and on the classroom teacher. We now try to separate these effects. Table 3 shows results broken down by room and treatment, with individual human tutors identified by two-letter codes. Why did some treatment groups do better? That is, to what extent can we assign credit for outcome differences between treatment groups to treatment, rather than to teacher effects, and/or to interactions between teacher and treatment, such as teacher cooperation with tutors?

<<insert near here: Table 3: Pretest and gain on each measure, by grade, classroom, and treatment group >>

To address this question, we expanded our ANOVA model to include classroom as a factor, and its interaction with treatment (updating the significant set of covariates for each outcome measure accordingly). The classroom variable encoded the identity of each student's classroom in order to model teacher effects. Including this variable in the model should expose differences between treatment groups that are actually due to teacher effects. Conversely, it may also reveal treatment differences previously masked by teacher effects.

In accordance with recent practice in educational statistics, we treated classroom as a random effect, and treatment as a fixed effect. This practice acknowledges teacher and social effects that cause the performance of different students in the same classroom to be correlated rather than independent. It models teachers as drawn from some distribution. We wish to draw inferences that will generalize to other teachers from this distribution, not just to future classes of the specific teachers in the study.

Which factors were significant in the expanded mixed effects model? In grade 2, neither treatment nor class was significant as a main effect for any outcome variable. Their interaction was significant for Word Attack (p=.025 with Word Attack and Word Identification pretest scores as covariates), almost significant for Word Comprehension (p=.054, with no significant covariates), and suggestive for Passage Comprehension (p=.103 with Word Comprehension and Passage Comprehension pretest scores as covariates). In grade 3, treatment was significant as a main effect for Word Attack (p=.016, with no significant covariates) and a main effect trend for Passage Comprehension (p=.086 with Word Comprehension and Passage Comprehension pretest scores as covariates; p=.027 with just Passage Comprehension). Treatment-class interaction was suggestive for Word Comprehension (p=.150 with Word Comprehension and Passage Comprehension pretest scores as covariates; p=.075 with just Word Comprehension). No other main effects were significant or even suggestive (p<.1). We did not attempt further analysis of interactions where there were no main effects, such as for Word Comprehension, because they tell us merely that some treatments worked better than others in certain specific classrooms.

To identify differences in effectiveness between two treatments, we ran mixed effects contrasts using the same covariates as before. Unlike SPSS's standard pairwise comparison or our earlier 1-way ANOVA, both of which identify significant differences between treatment *groups*, this procedure identifies significant differences between *treatments*, controlling for class effects – to the extent possible. Each class had students in at most two treatment groups, so we used Type IV sum of squares to cope with the resulting missing cells, but the results were the same as with the more commonly used Type III. Without Bonferroni correction for multiple comparisons, this

procedure found treatment effects for human tutoring over the Reading Tutor in grade 3 Word Attack (p=0.037), and for human tutoring over the control condition in grade 3 Passage Comprehension (p=0.058). Pooling human and automated tutoring yielded a significant main effect for tutoring on grade 3 Passage Comprehension (p=0.006).

How should we interpret these findings? The third graders who used the Reading Tutor outgained the baseline group in Word Comprehension and Passage Comprehension – but why? Did they just happen to have better teachers? After all, adding classroom to the model rendered insignificant the treatment effect for grade 3 Word Comprehension. However, it *strengthened* the main effect of treatment for grade 3 Passage Comprehension. Moreover, the mixed effects model showed no main effects for classroom in either grade on any subtest. We cannot conclude from the data that superior gains were due to teacher effects, but neither can we conclusively exclude this possibility, except for the human tutor group.

This ambiguity of attribution stems from study design and sample size. The study design was a hybrid of between- and within-class designs. Comparisons between human tutors and baseline were almost entirely within-class, thereby controlling for differences among teachers. However, comparisons of the Reading Tutor to the human tutor and baseline groups were entirely or almost entirely between-class. To rule out teacher effects, a between-class design would need many more than 6 classes per grade, ideally with random assignment of class to condition.

We can try to sharpen the evaluation of human tutoring by restricting it to a paired comparison that exploits the stratified random assignment to treatment. Recall that tutored and baseline students were matched by pretest scores within class. That is, we ordered the 12 study participants in each class by total reading score on the WRMT, and paired them up accordingly: the bottom two, then the next lowest two, and so forth. Then we randomly assigned one student in each pair to the human tutor condition, and the other one to the baseline condition. Consequently the difference between their gains reflects the effect of tutoring, since they had comparable pretest scores and the same classroom teacher.

Accordingly, to compare the two students in each intact pair, we defined outcome variables for the differences between their (actual) gains, and used a 2-tailed T-test to compare these differences against zero. For the 26 intact pairs as a whole, no differences were significant. When we disaggregated by grade, we found no significant differences in grade 2 (n=14 intact pairs), and trends favoring human tutoring in grade 3 (n=12 intact pairs) for Word Comprehension (p=0.085) and possibly Word Identification (p=0.118), but not Passage Comprehension (p=0.364). Why not? One possibility is that the increased sensitivity of the paired analysis was not enough to make up for the reduction in sample size caused by excluding unpaired students. Another possibility is that pairing students did not control for differences in pretest scores as effectively as treating them as covariates in the mixed effects ANCOVA.

Did individual tutors differ? That is, were human tutors equally effective, or did any human tutors differ significantly in the impact they contributed over and above classroom instruction? Recall that each human tutor was assigned to a different classroom, as shown in Table 3. Control group gains differed as much as 12 points from one classroom to another (for example, Word Comprehension in rooms 305 versus 309), not counting rooms with only 2 or 3 students in the control group. In general teacher effects might explain such outcome differences more parsimoniously than differences between tutors. How can we tease apart tutor effects from teacher effects?

To deal with this confound between teacher and tutor, we constructed an ANCOVA model to predict gain differences from blockmate, with pretests as covariates, and looked for main effects of classroom. This model already controls for teacher effects by subtracting gains made by students in the baseline condition in the same room. Therefore any main effects of classroom should indicate differences among individual tutors in their impact over and above classroom instruction in their respective classrooms.

Looking at gain differences between human tutored students and their matched classmates in the baseline condition, we found a suggestive (p=0.068) main effect of tutor on second graders' Word Identification. As Table 3 shows, ME's students outgained the baseline students in room 208; the gains difference, adjusted to control for significant pretest covariates, was 4.20 ± 6.53 SD. In contrast, MB and AC's students gained *less* than their matched classmates in the baseline condition in rooms 205 and 209, with an adjusted gains difference of -2.77 ± 7.22 SD. If we measure a tutor's impact by comparing her students' (adjusted) gains against those of their matched classmates in the control condition, then ME had significantly higher impact in Word Identification than the other two second grade tutors (p=0.019 without Bonferroni correction).

While ME helped students on Word Identification more than the other tutors, ME's students gained the least with respect to paired classmates on Word Comprehension (-13.40 ± 12.46 versus -1.00 ± 8.19 and 5.33 ± 12.56). The analysis of gain differences yielded suggestive but inconclusive results (p=0.111). However, an analysis of normed score gains showed that students tutored by MB in room 205 gained significantly more in Word Comprehension (9.81 ± standard error 2.48) than those tutored by ME in room 208 (-2.22 ± standard error 2.52).

In cases where tutored students gained significantly more in one room than in another, should we credit their tutor – or their classroom teacher? To answer, we examine the mean gains of the tutored and baseline groups in both rooms. The baseline students in room 205 gained somewhat *less* (4.33 ± 10.82) than those in room 208 (7.83 ± 6.31). So tutor MB in room 205 had unambiguously more impact on Word Comprehension gains than tutor ME in room 208.

We also checked for teacher effects in classrooms that used the Reading Tutor. Those rooms did not have enough students in the baseline condition to allow within-classroom comparisons. Instead, we compared mean pretest scores and gains of students who used the Reading Tutor in different classrooms. In second grade, we found no significant classroom gain differences within the Reading Tutor group. But in third grade, we found that students who used the Reading Tutor in room 303 gained significantly or suggestively less on four out of five measures than students who used the Reading Tutor in two other third grade classrooms, even though room 303 had the median pretest score of those three classrooms on all five measures. Room 303 gained less (p=0.001) on Word Attack than rooms 301 and 304 (-9.20 ± 6.78 versus 0.53 ± 6.27), less (p=0.037) on Word Identification than room 301 (-1.20 ± 3.65 versus 4.88 ± 4.67), less (p=0.103) on Word Comprehension than room 304 (2.21 ± 6.38 versus 6.02 ± 7.23), and less (p=0.088) on fluency than rooms 301 and 304 (17.79 ± 13.01 versus 27.44 ± 5.60). Might room 303's consistently lower performance be due to a difference in how – or how much – its students used the Reading Tutor?

In cases where tutors differed significantly in impact, it may be informative to compare their tutoring processes. Accordingly, we will revisit these outcome differences later to see if they might be explained by differences in process variables. But first we describe those process variables and relate them to outcomes. First we compare human and automated tutoring, based on videotaped sample sessions of each. Next we compare variables measured for all the tutoring sessions. Finally we relate process variables to outcomes.

Micro-Analysis of Student and Tutor Behaviors in Videotaped Sessions

We now turn our attention from the outcomes to the processes that produced them. Overall, what were the tutoring sessions like? How were human and automated tutoring similar? How did they differ? To answer these questions, we videotaped, coded, and analyzed 40 of the 6,080 human and computer tutoring sessions over the course of the year. To make this small sample capture some of the variation among students and tutors, we tried to include sessions of the Reading Tutor and the different human tutors helping students of low, medium, and high reading ability in each grade relative to the study sample, based on their pretest scores.

While the top-level activities of assisted reading and journal writing were common to the two tutoring environments, exactly how the tutoring experience plays out in the two environments could vary substantially. The

Reading Tutor's behavior is algorithmically defined, but is the only generally predictable component in the sessions. The human tutors exercised substantial latitude in the support they provided. Students may vary in help seeking behavior, or even in task engagement more generally. To compare the students' learning experience at a more detailed level in the human tutor and Reading Tutor sessions, 20 sessions of each type were videotaped and coded. In the Reading Tutor condition, eight sessions with second grade students and twelve sessions with third grade students were videotaped. In the Human Tutor condition, seven sessions with second grade students and thirteen sessions with third grade students were videotaped. All the human tutors are represented in the sample.

Session duration: Tutoring sessions were scheduled to last 20 minutes, but part of this time was devoted both to starting up and to finishing up activities. Table 4 displays the actual work time for the sessions, defined as the elapsed time beginning with the presentation of the first reading, writing, or vocabulary stimulus and ending with removal of the last such stimulus at the conclusion of the session. Average effective working time was similar across the four conditions, ranging from a low of 14.2 minutes in the third grade human tutor condition to a high of 18.8 minutes in the third grade Reading Tutor condition. Thus both tutoring treatments apparently spent comparable time per session on assisted reading and writing.

<<insert near here: Table 4: Mean Session Times (minutes) and Mean Reading Rate (text words per minute)>>

It is interesting to note that even though logging in could take half a minute, the human tutors actually averaged more time than the Reading Tutor in our 40-session sample on non-work activities such as getting started or awarding stars at the end of a session. However, we cannot reliably generalize this difference to the six thousand sessions that weren't videotaped, because duration and work time are exactly the sort of thing that might be influenced by observation, and we don't know their values for the sessions we didn't videotape. The reasons are different but interesting. The human tutor logs listed sessions as occurring exactly every 20 minutes, suggesting that the human tutors may have entered session times in advance rather than recording the true start and end times. As for the Reading Tutor, session duration and work time might in principle be computed from its detailed logs, but in practice it was infeasible to do so, both because their format was not designed for that purpose, and because it is surprisingly hard to define session duration to deal with such phenomena as failed login attempts and sessions interrupted by the Reading Tutor timing out (Mostow, Aist *et al.*, 2002; Mostow, Aist *et al.*, in press).

Waiting time: In viewing the videotapes it is clear that students were generally on-task in both the Reading Tutor and human tutor conditions when stimuli were present. (Of necessity, we can't be certain this conclusion generalizes to sessions that were not videotaped – especially in the Reading Tutor condition, where students were not individually supervised). However, it also became apparent in viewing the videotapes that students in the Reading Tutor condition spent appreciable time waiting for the computer to respond to their reading, writing, and vocabulary performance (e.g., to present a new stimulus sentence after the student read the current sentence).

Table 4 includes students' mean waiting time in the Reading Tutor condition. Waiting time is defined as the time during which there is neither a visual stimulus on the screen for the student to process, nor an auditory stimulus being presented. Waiting time includes time the student spent rereading the sentence when the Reading Tutor did not respond fast enough to the first reading, because it was waiting to make sure the student was done speaking. Another source of waiting time was time the Reading Tutor spent in preparatory computation before displaying the next sentence or other stimulus. Off-talk conversation with the teacher or other students was rare in the videotaped Reading Tutor sessions, and would not count as waiting time unless the student was waiting for the Reading Tutor to generate a task. However, off-task time often occurred when a student looked away from the screen while waiting, thereby failing to notice a newly displayed stimulus at first.

Waiting time accounted for approximately 45% of total session duration. This waiting time was not necessarily wasted, since students might have been thinking about the task as they waited, for example reflecting on what they had just read. However, it may be possible to increase the Reading Tutor's impact by decreasing this time. The

unexpectedly large fraction of time spent waiting for the Reading Tutor to respond led us to modify later versions of the Reading Tutor to respond more promptly.

Assisted reading rate: The remaining analyses focus on assisted reading. (In two of the videotaped human tutor sessions – one with a second grader, and one with a third grader – the student did no reading at all, only writing, so the sample size for these conditions is decreased by one in the subsequent analyses.) Table 4 displays students' mean assisted reading rates, defined as text words per minute (but not necessarily read correctly, as discussed below). Two measures of reading time are used to compute reading rate in the Reading Tutor condition. One measure employs total elapsed reading time, including waiting. The second measure, "net" reading time, excludes waiting time when there were no novel words on the screen to be read.

Not surprisingly, reading rate as a function of total elapsed waiting time was slower in the Reading Tutor condition than in the human tutor condition. Reading rate in the second grade Reading Tutor sessions was about 70% of reading rate for the human tutor sessions. In the third grade, reading rate in the Reading Tutor sessions was only about 40% of reading rate in the human tutor sessions. However, if we exclude waiting time and compute reading rate when there were words available to read, overall reading rates across the two tutor conditions were quite similar. In the second grade, net reading rate was actually about 40% faster in the Reading Tutor condition, while in the third grade reading net rate was about 20% slower in the Reading Tutor condition.

Errors and help requests: Students could invite reading assistance from a tutor either by making an error in reading or by asking for help. Reading errors included omitted words, inserted words, and substitutions (both other words and non-words). Students in the Reading Tutor condition could explicitly request help on an individual word by clicking on the word, or request help on an entire sentence by clicking below it. Students in the human tutor condition were never observed to ask for help in the videotapes, but pronounced pauses in reading were interpreted by the tutors as implicit help requests and elicited word-level help. The top section of Table 5 displays the frequency of students' reading errors, word-level help requests, and sentence-level help requests per session. (Students in the Reading Tutor condition sometimes clicked more than once on a given word, receiving a different form of word-level help for each click. Only the initial click on a word is reflected in this count.) Raw error frequency per session was similar across grade and treatment. However, the frequency of help requests per session in the Reading Tutor condition was 3-4 times the frequency in the human tutor condition.

<<insert near here: Table 5: Reading Errors and Help Requests>>

Student reading errors and help requests represent an objective, approximate measure of student learning opportunities. (This is an approximate measure since students may guess correctly on words they don't really know, and may stumble and/or ask for help on words they do know.) The raw frequency is a reasonable estimate of the rate at which such opportunities arose in this study, but this measure is subject to variation in how much time students actually spent reading in the tutoring session. The rate at which errors and help requests occurred per text word processed is a more general measure of how often these opportunities arose. The lower section of Table 5 displays the rate at which errors occurred per word processed, if and how these errors were corrected, and the rate per word processed at which the students asked for help. The bottom half of the table also distinguishes in the Reading Tutor condition between stories the students chose for themselves and stories the Reading Tutor chose.

An important observation emerges from the error rates in the lower half of the table. To put these error rates in perspective, a reader's frustration level is traditionally defined as more than 1 error per 10 words, and a reader's instructional level as about 1 error per 20 words (Betts, 1946). However, the availability of immediate assistance in an individual tutoring situation should allow more challenging text – especially in the Reading Tutor, which gives unlimited help on individual words or entire sentences. It is therefore interesting to compare how often students made errors and requested help.

The Reading Tutor and the human tutors selected equally challenging stories; the error rates were similar in these conditions. In contrast, students chose substantially easier stories, at least in the second grade. Students' average

error rate was substantially lower on stories they chose for themselves, whether because the stories were at a lower level, or because they were rereading them. Second grade students made only about 1 error per 50 words processed on stories they selected themselves in the Reading Tutor condition, but 1 error per 11-12 words on stories selected by the Reading Tutor or a human tutor. In the third grade, the error rates were more similar among the three conditions.

Table 5 shows the disposition of errors, the percent corrected by the tutor, the percent self-corrected by the student, the percent on which the student asked for help (which overlaps with the other categories), and finally the percent uncorrected. Table 5 reveals a potentially important difference between the Reading Tutor and human tutor conditions. On average across grades, almost 90% of errors in the human tutor condition were corrected. About 75% were corrected by the tutor and about 15% by the students. In the Reading Tutor condition, fewer student errors were corrected. The percent of corrected errors was similar across the two grades, but varied with story selection. Almost 80% of errors were corrected in student-selected stories, versus only 50% in stories selected by the Reading Tutor.

Table 5 includes an approximate breakdown of tutor corrections into explicit and incidental. All of the human tutor corrections were explicitly focussed on specific misread words. In contrast, many of the Reading Tutor "corrections" were not explicitly focused on individual words, but incidental in the course of reading the entire sentence. Often the Reading Tutor would fail to detect a miscue, but would read the sentence aloud anyway, either because it (rightly or wrongly) detected a miscue elsewhere in the sentence, or because the student was reading so haltingly as to put comprehension in doubt. The impact of such implicit corrections depends on whether students notice them. If not, then the Reading Tutor left even more errors *effectively* uncorrected.

Finally, the bottom half of Table 5 also displays the rate of help requests per text word processed. Note that students in the Reading Tutor condition requested word-level help at just about the same absolute rate that they were making reading errors, regardless of whether the student or tutor chose the story. The rate at which human-tutored students asked for help (defined as a pronounced pause in reading) was much lower than the rate at which they made errors. This data answers two questions raised above in the discussion of difficulty levels for assisted reading. First, students were likelier to request help in the Reading Tutor condition. Second, they read text that without such help would have exceeded their frustration level. That is, students made at least one error or help request per 10 words in stories selected by the Reading Tutor.

We examined the Reading Tutor videotapes for evidence of sentence level "help abuse" by students. In its most extreme form, a student might ask the Reading Tutor to read each sentence aloud as it appeared before attempting to read it himself. Among the 20 videotaped sessions, we found one second-grade student who asked the Tutor to read half of all the sentences when they were first presented. In a second second-grade session, a different student asked the tutor to read a quarter of all the sentences when they first appeared. Among the other 18 sessions there was no evidence of such abuse. 13 students never asked the Tutor to read a whole sentence.

Likewise, one student at the third-grade level asked for help on 19% of the individual words in the story, one second-grade student asked for help on 23% of individual words, and another second-grade student asked for help on 17% of words. The remaining 17 students asked for help on less than 10% of words. Taken together, these findings are consistent with the possibility that relatively few students over-used help – at least when being videotaped.

Tutor interventions: Neither the Reading Tutor's nor the human tutors' interventions were limited to reading assistance opportunities. We distinguish three categories of reading assistance. *Pre-emptive reading assistance* gave advance help on how to pronounce words before the student reached them. *Reading assistance opportunities* consisted of responses to reading errors and help requests as described above. *False alarms* were interventions after the student read a portion of text with no apparent errors. We distinguish two additional categories of tutor interventions related to reading. *Praise and backchanneling* were tutor utterances that praised the student's

performance, confirmed the student's performance, and/or encouraged the student to continue. *Discussion of meaning* discussed the meaning of a word or text after it was read. We exclude other categories of tutor interventions, such as prompts about how to operate the Reading Tutor, e.g., "you can click on a word when it has a box around it."

Table 6 displays both the rate of tutor intervention per text word processed and the percent of overall tutor interventions accounted for by each category. The first row displays overall tutor intervention rate, defined as the mean number of interventions per word of text, counting multiple interventions on the same word as separate interventions. The Reading Tutor intervention rate averaged about 0.2 (1 intervention for every 5 words processed). This overall rate was about double the average intervention rate for the human tutors.

<<insert near here: Table 6: Categories of Tutor Intervention>>

The middle section of the table summarizes the three categories of Reading Assistance: Pre-emptive, Response to Errors and Help Requests, and False Alarms. Note that the total intervention rate across these three reading assistance categories was higher for the Reading Tutor than for the human tutors. Also, there was a striking difference between the Reading Tutor and human tutors in the distribution of interventions among the three reading assistance subcategories. The human tutor interventions all focused on student reading errors and help requests, while the Reading Tutor's interventions split more evenly among the three subcategories.

As the bottom of the table shows, the percentage of praise, confirmation, and backchanneling was very similar for the Reading Tutor and the human tutors. These responses were essentially meta-utterances designed to encourage student progress in reading the text. Praise utterances complimented the student, e.g., "outstanding," "super work," "good." Confirmation utterances signalled that the student had performed correctly "okay," "good," or repeated a word the student had read. Backchanneling consisted of non-verbal utterances (e.g., "mm-hmm," "uh-huh," "hmmm," coughing) designed to catch the student's attention if necessary or signal that the student should continue.

It is difficult to draw a sharp distinction among these meta-utterances. For example, "good" could be either praise or confirmation. Similarly, "mm-hmm" could be either confirmation or backchanneling. The human tutors and Reading Tutor emitted these flow-control utterances at about the same rate, but with different meaning – sometimes even for the same utterance. For example, a human tutor might sometimes say "mm-hmm" to confirm that the student had read correctly. In contrast, the Reading Tutor said "mm-hmm" only to encourage the student to continue reading, because the limited accuracy of speech recognition precluded certainty as to whether the student had read correctly (Mostow & Aist, 1999b).

About 8% of human tutor interventions either engaged the student in a discussion of passage meaning or discussed the meaning of a word after the student (or tutor) pronounced it correctly. The Reading Tutor did not engage in this behavior. One reason is that speech recognition technology was not yet accurate enough to support spontaneous discussion. The Reading Tutor did provide occasional vocabulary assistance by inserting "factoids" into the text.

<<insert near here: Table 7: Categories of Reading Assistance>>

Types of reading assistance: Table 7 summarizes the different types of reading assistance offered by the tutors. This table collapses across all the situations in which tutors gave help, namely pre-emptive assistance, responses to student errors and help requests, and false alarms. To compare how human tutors and the Reading Tutor gave help, we classified assistance into several types and analyzed their relative frequency. Some tutor interventions *focused the student's attention* on a word without providing any pronunciation scaffolding. In the Reading Tutor these interventions included underlining the word, backchanneling, and recuing (reading a sentence up to, but not including the focus word). For human tutors this category included recuing and exhortations essentially to try

harder, e.g., "look at the word", "think about the letters." Sometimes the tutor *read a word aloud*. Sometimes the tutor *read an entire sentence aloud*. *Sounding out a word* included three types of exaggerated pronunciations, which emphasized the syllable structure of a word, the onset/rime structure of a word, or the pronunciation of its individual phonemes. Sometimes the tutor called the student's attention to a *rhyming word*. Some interventions *focused on letter-sound correspondence* by discussing specific letters in the word and how those letters are pronounced, but did not necessarily discuss the generality of the correspondence and did not cite a rule name. In contrast, human tutors occasionally *cited a letter-sound pattern rule*, either "Magic E" (when a vowel is followed by a consonant and e, pronounce it as a long vowel) or "Two vowels walking" (when two vowels occur in succession, just pronounce the long version of the first vowel). *Spelling a word* either told the student to name or engaged the student in naming the letters in a word. Occasionally, human tutors *gave a semantic cue* to help a student pronounce a word in his or her spoken vocabulary (e.g., "it's an orange vegetable" for carrot). We distinguish this type of help from discussing the meaning of a word after it has been pronounced.

Note that a tutor could provide more than one type of assistance when the student misread a word or clicked for help on a word. For instance, recuing a word might be sufficient if the student had merely slipped or misapplied word attack skills. But if recuing failed, a human tutor would offer one or more additional types of help until the word was pronounced correctly. In the Reading Tutor, the student could click for help repeatedly on the same word and get different types of assistance. Some students would click on a word until the Reading Tutor spoke it – sometimes not even waiting to hear the entire hint before clicking again, which caused the Reading Tutor to interrupt itself. Table 7 tallies every instance of tutor assistance, including multiple instances on a given word, whether or not the assistance was completed.

The Reading Tutor and human tutors displayed similar percentages of assistance responses in two categories: focusing on a word, and exaggerated sounding out of words. At the second grade level, there was a pronounced difference between the Reading Tutor and human tutors. The human tutors were far more likely to provide letter-related assistance (letter-sound correspondence, sound pattern rule, or spelling). Almost 40% of human tutor assistance was letter-related, while only 5% of Reading Tutor assistance was letter-related. In contrast, the Reading Tutor was far more likely than the human tutors to read either the single word on which the student needed assistance, or a full sentence including such a word. Just over 50% of Reading Tutor responses consisted of reading a word or sentence aloud, versus only 18% of human tutor responses. At the third grade level, the rate of these reading-aloud responses was more similar (about 55% for the Reading Tutor and 46% for human tutors), as was the rate of letter-related responses (just over 5% for the Reading tutor and 12% for the human tutors).

It is interesting to relate these findings to previous studies. A study of talking-computer assistance on demand for first and second graders (Wise, 1992) found that "presenting words as wholes is at least as helpful for short-term learning as presenting them segmented," but (Wise & Olson, 1992) found that "for children 10 years or older, training with intermediate speech feedback led to greater benefits in phonological coding skills than training with word-only feedback."

Didactic versus interactive assistance. The same content can be conveyed in different ways. Didactic assistance conveys content by telling it. Interactive assistance conveys content by engaging the student in helping to construct it. For example, a word can be sounded out didactically by the tutor, or interactively by getting the student to sound out the word.

The Reading Tutor's assistance conveyed any content didactically. To constrain its speech recognition task, the Reading Tutor was designed to avoid eliciting any speech from the student other than reading the current sentence aloud. It lacked the speech understanding capabilities required to engage in other forms of spoken dialogue, such as cooperative efforts to sound out a word.

In contrast, human tutors could and did engage in such dialogue. When offering word attack support (exaggerated sound out, rhyme, letter-sound correspondence, sound rule pattern), second grade tutors engaged students

interactively 56% of the time and presented the information to students didactically 44% of the time. Third grade tutors engaged students interactively 62% of the time and presented the information didactically 38% of the time. More generally, excluding only the "focus on word" category, second grade tutors interactively engaged students 46% of the time in presenting corrective information and didactically presented the information to students 54% of the time. Third grade tutors interactively engaged students 27% of the time and didactically presented the information to students 73% of the time.

This contrast between the Reading Tutor and human tutors is important. Students may learn information better by helping construct it themselves than simply by being told. However, it should be emphasized that this contrast applies only to the specific issue of how the information content of reading assistance was conveyed, and not to the nature of the tutorial dialogue in general. The Reading Tutor was highly interactive in the sense of eliciting and responding to the student's oral reading. Much of its assistance gave the student only a hint about how to read a word. This assistance was didactic only in the narrow sense that the Reading Tutor conveyed the information content *in the hint* by telling it, rather than by engaging the student in an interactive process of constructing it.

"Pause the Video" experiment: To evaluate how appropriately the Reading Tutor chose which responses to employ, we tried using a panel-of-judges methodology for evaluating expert systems. Three professional elementary educators watched 15 video clips of the Reading Tutor listening to second and third graders read aloud, recorded so as to show both the Reading Tutor and the student's face reflected in a mirror. Each judge chose which of 10 interventions to make in each situation. To keep the Reading Tutor's choice from influencing the expert, we paused each video clip just before the Reading Tutor intervened. After the judge responded, we played back what the Reading Tutor had actually done. The judge then rated its intervention compared to hers. We only summarize this experiment here; for details, see (Mostow, Huang, & Tobin, 2001).

For example, in one such clip the text word was "look," and the student said "foot … lo… lo…" After seeing this portion of the video clip, the judge selected the intervention she would have chosen, such as *Sound Out*: "/l/ /oo/ /k/." Then the experimenter played back what the Reading Tutor actually did, in this case *Rhymes With*: "rhymes with shook." The judge then rated this choice, compared to her own, as "better," "equally good," "worse but OK," or "inappropriate."

Although the judges seldom agreed on what specific intervention to use, they generally chose from the same four interventions. Sounding out (either phoneme by phoneme or syllable by syllable), reading a word aloud, or rhyming it with another word accounted for 76% of the judges' responses, and 14 of the actual Reading Tutor responses in the 15 video clips. The judges rated the Reading Tutor's choices as better than their own in 5% of the examples, equally good in 36%, worse but OK in 41%, and inappropriate in only 19%. The lack of more specific agreement and the surprisingly favorable ratings together suggest that either the Reading Tutor's choices were better than we thought, the judges knew less than we hoped, or the clips showed less context than they should.

Analysis of Process Variables from Comprehensive Records

We now examine process variables we measured to characterize possible differences between the two tutoring conditions. These measurements were motivated by the expectation that the human tutors would far surpass the Reading Tutor. We hoped that identifying differences in process variables might help us explain outcome differences and improve the Reading Tutor.

Our micro-analysis was based on sample videotaped sessions that may or may not be representative of other students or tutoring sessions. In contrast, we now examine process variables based on *comprehensive* records at various levels of detail for all 6,080 tutoring sessions.

We omitted classroom instruction from this comparison. The process variables were not feasible to measure for the classroom instruction received by each individual student. The information in our teacher questionnaires was

not adequate to estimate even approximate means for each classroom, which would have required systematic and detailed classroom observation (Foorman, Francis *et al.*, 1998). In fact these variables may not even be well-defined for the students in the baseline condition. For example, how many tutoring sessions did baseline students not have, and when did they not have them?

<<insert near here: Table 8: Comparison of process variables for Reading Tutor (RT) and human tutoring (HT), by grade>>

Table 8 compares Reading Tutor and human tutor process variables. The symbols << and >> indicate significant differences between condition, and the symbol <? indicates a trend. The annotations + and ++ (or - and --) identify the significantly highest (or lowest) of the values for a given treatment in 3 classrooms from the same grade. For example, students used the Reading Tutor for significantly more days in room 201 than in rooms 211 and 212. The annotations ++ and -- indicate that the difference from one or both of the other two rooms is reliable ($p < .01$), while + and - indicate that the difference is suggestive ($p <\sim .1$). Comparisons between and within treatment groups reveal some interesting similarities and differences.

Session frequency: Treatment fidelity was considerably better than in the spring 1998 study. Frequency of tutoring approached the target of having daily sessions. Session frequency varied significantly among rooms in the Reading Tutor condition but not in the human tutor condition.

Why? Reading Tutor usage depended on classroom teachers' cooperation as "gatekeepers" to put their students on the Reading Tutor. In contrast, human tutors bore responsibility for tutoring their assigned students, whom they could come get if necessary. Reading Tutor usage also relied on teachers for frontline assistance in fixing technical glitches. For example, the Reading Tutor often needed to be relaunched, sometimes more than once a day. The Reading Tutor rebooted automatically every night – but waiting until then would prevent usage for the rest of the day. Recovery was faster if the teacher (or students) did it. Likewise, headsets occasionally broke and had to be replaced. Our Educational Research Field Coordinator visited the school as often as 3 times a week – but waiting until then instead of plugging in a spare headset would prevent usage in the interim.

Despite its limitations, students who used the Reading Tutor averaged almost as many sessions as human-tutored students in grade 2, and significantly more in grade 3. In both grades, the classroom with the highest Reading Tutor usage averaged about 90 days on the Reading Tutor – 20 or 30 more than the human tutors. Thus teachers who were willing and able to cope with the Reading Tutor's limitations succeeded in making it considerably more available than the human tutors. This contrast is even more dramatic in view of the fact that each human tutor worked with 6 students, while each Reading Tutor served 10 (and in two rooms, occasionally even more). The human tutors were assigned to other duties in the morning, and were available to tutor only in the afternoon. Although human tutors were not subject to frequent breakdowns, they were occasionally absent or pulled off to substitute-teach or attend professional development or other events. The human tutors logged absences but almost no truncated sessions. In contrast, the Reading Tutors stayed in their classrooms all day (with rare exceptions for repairs), on dedicated computers. Aside from technical problems, Reading Tutor usage was limited primarily by the classroom schedule.

Words read: The number of words read per session differed considerably between treatments. This difference appeared to favor human tutors, but that may be an artifact of the following difference in accounting. The human tutor word counts, calculated from tutor logs, include partial story readings. The Reading Tutor word counts, computed from student portfolios, include only stories the student finished reading. Although the portfolios record which stories the student started without finishing, they do not show how much or little of them the student read.

Story difficulty: Controlling for materials by having the human tutors use hardcopies of the same set of stories as the Reading Tutor facilitated comparison of story level between conditions. Stories that students finished reading in the Reading Tutor averaged half a level easier than with human tutors (1.1 vs. 1.8 in Grade 2, and 1.7 vs. 2.2 in

Grade 3). The Reading Tutor chose stories at the same average level (1.8) as the human tutors in Grade 2, and slightly harder (2.5) in Grade 3. But students chose easier stories on their turn. They also reread stories nearly twice as often as human tutors permitted. Rereading an old story is easier than reading a new story.

Writing: The human tutors and the Reading Tutor included writing as well as reading. In the 1999-2000 Reading Tutor, students could write (and optionally narrate) free-form stories, and edit stories they had written previously. Human tutors employed more varied writing activities. Our data coder categorized 53% of them as writing an original story, 22% as spelling or punctuation practice, 6% as copying a story directly from the reading, 4% as a composite of two or more of the above categories, and 7% as none of the above, e.g., word practice, questions, or coloring. Human tutoring sessions included writing almost twice as often as Reading Tutor sessions. However, this figure does not show the relative amounts of time spent on writing. Some students chose to spend considerable time writing stories in the Reading Tutor, while others spent none. In contrast, human tutors may have spent a regular but small part of each session on writing. However, we did not ask the human tutors to log the duration of each activity within a session, lest we overload their tutoring with bookkeeping.

Relation of outcomes to process variables

To relate outcomes to process variables, we correlated them against each other, and against process variables. Typically, the lower a student's pretest scores, the greater the student's gains in normed scores, whether because of regression to the mean, or because instruction helped poorest readers the most. Partial correlations factor out this effect by accounting for the residual variance in gains after controlling for covariates. Table 9 shows the resulting partial correlations, controlling for the same significant covariate pretest scores listed earlier in Table 2. The partial correlation of two gains is symmetric only when they have the same covariates. For example, Word Identification and Word Comprehension gains have the same covariates, so the third cell in the second row of Table 9 (with values –.091 and .337) mirrors the second cell of the third row. But Word Attack gains have different covariates, so the second cell in the first row (with values .059 and .314) does not mirror the first cell in the second row (with values .415 and .193).

We omitted classroom as a factor in order to avoid masking classroom effects mediated via or reflected in the process variables. For example, including Classroom as a factor would have obscured the effect of Sessions, because usage varied from one classroom to another. Consequently the models omit classroom effects not captured by the process variables, such as differences in teaching style (except insofar as they influence the process variables).

We correlated within rather than across conditions because the process variables were measured differently in the two tutoring conditions, thereby introducing potential systematic bias. As it turns out, the correlations varied considerably. Seldom was the same correlation significant in more than one grade or treatment condition. However, when a correlation was significant in one tutoring condition, the correlation in the other tutoring condition was generally consistent in sign, or else tiny in magnitude.

<<insert near here: Table 9: Partial correlations of gains with each other and with process variables, controlling for significant pretest covariates (**?**, *****, and ****** indicate respective significance levels of $p < .10$, $p < .05$, and $p < .01$)>>

How did gains correlate with process variables and other gains? Correlation does not imply causality. A process variable that predicts a gain might *reflect* rather than *explain* it. For example, story level may correlate well with fluency gains, but which is cause and which is effect? Nonetheless, the correlation is interesting either way: even if the process variable does not explain the gain, if it reliably predicts the gain it may be useful for automated assessment. We now summarize the significant partial correlations in Table 9 between process variables and gains. *Number of sessions* correlated with nothing in grade 2, and only with Word Attack gains for third graders who used the Reading Tutor. *Story level* correlated positively with gains in Word Comprehension, Passage

Comprehension, and fluency for both groups in grade 2, but only with fluency gains in grade 3. *Re-reading* correlated negatively with Word Comprehension gains for second graders with human tutors, though only suggestively, but positively and significantly for third graders who used the Reading Tutor, and with their Word Attack gains as well. *Writing* correlated negatively with Passage Comprehension gains for second graders with human tutors, and positively with fluency gains by third graders with human tutors, but both correlations were only suggestive. *Words read* correlated positively with gains across the board in second grade, significantly so for Word Identification and Word Comprehension in the Reading Tutor group, and for Passage Comprehension and Fluency in the human tutoring group. But in third grade, the number of words read correlated only suggestively with human-tutored students' Fluency gains, and not with any gains for the Reading Tutor group.

Relation to published correlations among WRMT subtests: Is Table 9 consistent with known correlations among WRMT subtests? (Woodcock, 1998) reports positive correlations among the four WRMT subtests for grade 3, ranging from .59 (between Word Attack and Passage Comprehension) to .72 (between Word Identification and Word Comprehension). Correlations for grade 1 manifest similar relationships but higher magnitudes. Correlations for grade 2 are not reported, but presumably lie somewhere between.

Correlations among students' scores on different subtests do not necessarily imply partial correlations among their *gains* on those subtests. Also, results reported here are for a population of students selected (by their teachers) as being in the bottom half of their respective classes. The resulting sample therefore differs from the norming sample used for the WRMT. Nevertheless, as the top of each half of Table 9 shows, most of the partial correlations among gains in this study were positive, including all the statistically significant ones. The negative partial correlations were not only insignificant but very weak – less than .1 in magnitude except for two correlations with Fluency, which is not one of the WRMT subtests. Thus the partial correlations among gains are qualitatively consistent with reported correlations among individual subtest scores.

Can process differences explain outcome differences between tutors? In particular, are the correlations in Table 9 consistent with significant outcome differences noted earlier between human tutors?

Tutor ME had significantly higher impact on Word Identification than tutors MB and AC, in the sense of being the only second grade tutor whose students outgained their matched classmates in the control condition. However, ME had the median of the three second grade tutors for every process variable except story level, which averaged only slightly lower than for tutor AC. Moreover, Table 9 shows that story level did not correlate significantly with Word Identification gains in grade 2; in fact, none of the process variables did. Moreover, tutor ME's students saw barely half as many words per session as tutor MB's. Thus the process variables we measured do not readily explain tutor ME's apparent impact on Word Identification gains.

What about tutor MB's second graders outgaining tutor ME's second graders in Word Comprehension? Their students averaged the same number of sessions (77). On average, MB's students read harder stories (level 2.8 vs. 1.2), consistent with the .378 correlation with Word Comprehension gains. MB's students reread fewer stories (11% vs. 21%), consistent with the suggestive -.448 correlation. MB's students wrote in fewer sessions (37% vs. 70%), consistent with the -.292 correlation. MB's students read more words (224 vs. 120), consistent with the .223 correlation. Thus the difference in Word Comprehension gains is consistent with these partial correlations. However, only one of them (with rereading) was stronger than .4 or even suggestively significant.

Two additional caveats are important here. First, MB's and ME's students accounted for about two thirds of the second graders from whose individual performance the correlations were derived, so consistency with the difference in their collective performance is hardly surprising. This consistency would be more impressive if the correlations were derived exclusively from other students. Second, correlations do not necessarily imply causality. Thus, we should view process variables correlated with positive outcomes merely as plausible explanatory candidates to investigate in future.

Can process differences explain outcome differences between classrooms? More specifically, what about classroom differences within the Reading Tutor condition? Recall that students who used the Reading Tutor in room 303 gained less in Word Attack, Word Identification, Word Comprehension, and Fluency than in rooms 301 and 304. Of course one possibility is simply that their classroom teacher was not as effective, which there were not enough baseline students to determine with any confidence. But another possible explanation is the fact that students in room 303 averaged significantly fewer sessions on the Reading Tutor – 57, compared to 70 in room 301 and 86 in room 304, as Table 8 shows. Sessions had a significant positive partial correlation with gains in Word Attack for third graders who used the Reading Tutor, as Table 9 shows. Room 303 averaged highest in story level, which correlated suggestively with fluency gains. Room 303 was lowest in rereading, which had a significant positive correlation with Word Attack and Word Comprehension gains. In short, process variables – especially lower usage, if it was due to teacher gatekeeping rather than student choice – might help explain why room 303 gained less than rooms 301 and 304 in Word Attack, Word Comprehension, and Fluency.

Can process differences explain outcome differences between treatments? That is, are the partial correlations consistent with differences between the Reading Tutor and human tutoring? Human tutoring significantly outgained the Reading Tutor only in Word Attack. The number of sessions and percent of rereading correlated positively and significantly with Word Attack gains by third graders who used the Reading Tutor. They may have gained from spending more time on the Reading Tutor, and by rereading old stories – but they didn't progress as much in Word Attack as their peers in other conditions. No other partial correlations were significant for either grade or treatment. Consequently the particular process variables we were able to instrument comprehensively shed little if any light on the difference in Word Attack gains across tutoring conditions.

We believe that differences between the Reading Tutor and human tutors in Word Attack are better explained by such factors as humans' superior hearing and consequent ability to detect oral reading miscues, as reflected in the percentage of miscues corrected in the videotaped sessions, and possibly by how they responded to miscues, for example with more frequent letter-related assistance in grade 2. We have identified several candidate explanations for why human tutors helped Word Attack more than the Reading Tutor did. These hypotheses are guiding our analyses of data from subsequent versions of the Reading Tutor, and our attempts to improve it.

1. *Students spent much of the time waiting for the Reading Tutor to respond*, and therefore read fewer words. Analysis of videotaped Reading Tutor and human tutor sessions showed that students read fewer words per minute in the RT than in HT. Analysis of tutor logs showed fewer total words read per session in the Reading Tutor than with human tutors. However, the latter comparison is subject to bias because it counts partially-read stories in the human tutor condition but not in the Reading Tutor, due to differences in how the two types of sessions were logged. Also, if differences in Word Attack gains were due solely to reading fewer words, we would expect to see similar differences in other subtests, especially Word Identification. But Word Identification gains did not differ significantly between conditions in either grade. In fact, Table 2 shows that the baseline group outgained the human tutor group in Grade 2 (though not significantly), and in grade 3 the Reading Tutor group gained at least as much in Word Identification as the human tutor group.

2. *Students requested much more help from the Reading Tutor than from human tutors*, and so got less practice decoding unfamiliar words independently. In particular, a few students over-used help in the Reading Tutor, so they got less practice because they made the Reading Tutor do too much of the work, either by making it read the entire sentence too often, or by clicking on a word repeatedly until the Reading Tutor spoke it.

3. *A much higher percentage of errors went uncorrected in the Reading Tutor than in human tutoring*, or at least were "corrected" only in passing, by reading the whole sentence. Putting corrective information into the environment is necessary but not sufficient – the student must notice it. Consequently students who used the Reading Tutor got fewer opportunities to practice correcting mistakes.

4. *The Reading Tutor spoke the word more often than human tutors*, so students got less practice decoding words based only on hints.

5. *The Reading Tutor gave letter-oriented help less often* (at least in grade 2), so students focused less on letter-sound mappings.

6. Even when it detected an error (or the student requested help), *the Reading Tutor did not engage the student in interactively constructing the pronunciation*, or gauge whether student was succeeding. Didactic interventions such as sounding out words for the student may have been less effective than helping the students sound out words themselves.

What about the outcome differences between the tutored and non-tutored students? Although we don't have process variables for the baseline group, we can still look at process-outcome correlations for possible clues. In particular, process variables that correlated with grade 3 gains in Word Comprehension and Passage Comprehension might suggest possible explanations for why those gains exceeded the baseline group. Table 9 shows that only one process variable correlated significantly with Word Comprehension or Passage Comprehension gains by either treatment group in third grade. *For the Reading Tutor group, the higher the percentage of rereading, the higher the gains in Word Comprehension* ($R=.433$, $p<.05$).

This finding surprised us. We expected Word Comprehension gains to *decrease* with the percentage of rereading, because the higher the percentage of new stories students read, the more new words they can encounter. In a separate analysis that counted just *distinct* words read in this study, (Aist, 2000, p. 5.5) found a positive relationship: "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, $p = .178$." Controlling for both Word Comprehension and Word Identification pretest scores strengthens this partial correlation to .26, $p = .189$.

The finding about rereading suggests that rereading can sometimes actually build vocabulary better than reading new text, by helping students understand new words they didn't grasp the first time around. This phenomenon would be consistent with research showing that younger children gain vocabulary from hearing repeated readings of the same story, and that rereading a passage can improve comprehension of it (Levy, Nicholls, & Kohen, 1993; NRP, 2000). It would imply that reading *only* new stories, which might expose the reader to new words, does not build vocabulary as well as rereading stories sometimes. For example, readers might learn more vocabulary by reading 50 stories twice than reading 100 stories once. Even though they would *see* fewer new words, the increased exposure might increase the total number of words they *learned*.

What's new here? There is already consensus that it generally takes multiple exposures to learn the meaning of a new word (Kamil, Mosenthal *et al.*, 2000, p. 270). However, typically these multiple exposures involve encountering the word in different contexts. The data suggest that more than one exposure to the *same* context might sometimes improve vocabulary *more than spending the same time reading new text*. However, caution is important here. Rereading correlated positively with Word Comprehension gains only for third graders who used the Reading Tutor, and did not correlate significantly with their Passage Comprehension gains. For human-tutored second graders, the correlation with Word Comprehension gains was actually negative (but only suggestive).

"Factoids" experiment to evaluate automated vocabulary assistance: (Aist, 2002a, b) models vocabulary growth as a product of new words encountered and the amount of learning from each encounter. Rereading may be one way to enhance that learning. Explaining new words is another. To test whether the Reading Tutor's vocabulary assistance was effective in explaining new words, we embedded an automated experiment to compare children's understanding of words the Reading Tutor explained, compared to words it did not. This experiment is reported in more detail in the journal version (Aist, 2001b) of a conference presentation (Aist, 2001a) based on a dissertation chapter (Aist, 2000). Here we summarize its design and results, and relate them to this study.

The experiment worked as follows. Just before displaying a sentence containing a new word, the Reading Tutor randomly decided whether to explain the word. If so, it inserted a short "factoid" relating the word to a (hopefully) more familiar synonym, antonym, or hypernym. For example, just before a sentence containing the word "astronaut," the Reading Tutor decided to explain it, that is, to assign the word "astronaut" to the experimental condition for this particular student. Accordingly, it displayed a factoid relating "astronaut" to a more familiar hypernym: "astronaut can be a kind of traveler". The reader read this factoid with the Reading Tutor's normal assistance. Then the Reading Tutor displayed the sentence "The Russians took the lead thirty three years ago by sending the first astronaut into space" and the reader resumed reading the story. The next day, the Reading Tutor asked the multiple-choice question "Which of these do YOU think means the most like astronaut?" with the randomly ordered choices "past master," "desperate," "best friend," and "traveler." If the randomized choice had assigned the word "astronaut" to the control condition, the Reading Tutor would have skipped the factoid and gone directly to the story sentence from the previous sentence, but would still have tested the word the next day.

Did inserting a factoid about a new word provide significant benefit above and beyond reading the word in context? Overall, no: In 3,359 randomized trials, students averaged 38% correct on the experimental (factoid) words, vs. 37% on the control words, and this difference was not statistically reliable. However, exploratory analysis showed that factoids *did* help significantly on rare, single-sense words (like "astronaut") tested 1-2 days later (44% vs. 26%, N = 189 trials), and suggested that factoids helped third graders more than second graders.

The factoid intervention probably explains at most a small part of third graders' advantage over the baseline in Word Comprehension gains, compared to the value of encountering many new words in context. Nonetheless, the factoids study demonstrated that an automated "invisible experiment" (Mostow & Aist, 2001) embedded in the Reading Tutor could test not only *whether* an individual tutorial intervention helped, but shed light on *when* it helped: which words and which students.

Conclusions

Until now, few studies have "directly examined the effects of using computer technology for reading instruction" (NRP, 2000, pp. 6-1, Ch. 6) – let alone over prolonged periods (Wise, Olson *et al.*, 1989; Wise, Ring, & Olson, 1999). Fewer still have compared computer technology to human tutors (Icabone & Hannaford, 1986). This study compared a daily 20-minute automated intervention – the 1999 version of Project LISTEN's Reading Tutor – to baseline classroom instruction and to one-on-one tutoring by certified teachers over the course of virtually an entire school year.

The biggest surprise was the similarity in outcomes among the three treatment groups. Thus the main result is that (except in Word Attack) the 1999 Reading Tutor yielded similar or greater gains than staying in classroom instruction, and rivalled one-on-one tutoring by certified teachers. The human-tutored group significantly outgained the computer-tutored group only in Word Attack. In grade 3, both the human- and computer-tutored groups outgained the classroom instruction group significantly in Word Comprehension and Passage Comprehension.

Effect sizes in this study were moderate to large, according to the criteria used by the National Reading Panel. "To judge the strength of an effect size, a value of 0.20 is considered small, 0.50 is moderate, and 0.80 is large" (NRP, 2000). These effect sizes are impressive given that the study manipulated only 20 minutes of treatment per day, compared to 1-2 hours of daily instruction in language arts. Over the entire study, students in the Reading Tutor and human tutor groups averaged only 20-30 hours of tutoring in total, depending on classroom.

The study design discriminated treatment effects from teacher effects better for human tutoring than for the Reading Tutor. One lesson is to prefer within-classroom comparisons when the number of classrooms is small.

The Word Comprehension results might be due to teacher effects, but the Passage Comprehension results were apparently due to tutoring, and are consistent with results from the within-classroom evaluation of the 1998 Reading Tutor (Mostow, Aist *et al.*, in press).

Besides informative micro-analysis based on 40 videotaped sessions, we analyzed process variables based on comprehensive records of *all* 6,080 tutoring sessions of the 92 students in the two tutoring conditions. These analyses revealed differences between the two tutoring conditions, differences between individual human tutors, and significant relationships between process and outcome variables. For example, one plausible reason for the difference in Word Attack gains is that the Reading Tutor provided explicit corrections for only half as many oral reading miscues as the human tutors, due to the limited accuracy of its speech recognition. Accordingly, we are working to improve its ability to detect miscues (Fogarty, Dabbish *et al.*, 2001; Mostow, Beck *et al.*, 2002). Significant outcome differences between individual tutors confirm the importance of decisions about which activities to work on (Aist & Mostow, in press; Juel, 1996). Session count, story level, words read, writing, and rereading were predictive of various gains – whether as cause, effect, or both. Rereading seemed to help third graders who used the Reading Tutor improve Word Comprehension more than reading only new stories. Analysis of the videotaped sessions also revealed contrasts in how tutors responded to miscues (Mostow, Huang, & Tobin, 2001). Automated experiments embedded in the Reading Tutor shed light on the effectiveness of its vocabulary assistance (Aist, 2001b). We are working to understand how specific automated interventions affect student learning (Mostow, Aist *et al.*, 2002; Mostow, Aist *et al.*, 2001), so that we can improve the Reading Tutor's effectiveness accordingly.

Acknowledgements

We thank the students and educators who participated in this research, and other members of Project LISTEN who contributed. This work was supported in part by the National Science Foundation under Grant Nos. REC-9720348 and REC-9979894, and by Greg Aist's National Science Foundation Graduate Fellowship and Harvey Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation or the official policies, either expressed or implied, of the sponsors or of the United States Government.

Tables

Table 1: Results of spring 1998 4-month within-classroom comparison

Subtest	Mean Pretest	Mean Pretest by Condition			Actual Gain			Significant Covariates	Gain Adjusted by Covariates			Main Effects p =	Effect Size
		Class	Commercial	RT	Class	Commercial	RT		Class	Commercial	RT		
All Grades		(n = 20)	(n = 21)	(n = 22)					(n = 20)	(n = 21)	(n = 22)		
Word Attack (normed)	85	85	83	86	5.6	7.1	5.6	WI, WA	6.1	6.5	5.3	1.00	
Word ID (normed)	83	84	82	83	1.2	2.9	2.2	WI	1.3	2.5	2.0	0.84	
Passage Comp (normed)	84	85	83	84	-1.9	0.4	**1.9**	PC	-1.9	0.4	**2.4**	**0.11**	**0.60**
Grade 2 (n=21)	83	86	82	81	-0.9	0.8	4.4	PC	-0.5	0.6	4.2	0.60	
Grade 4 (n=18)	80	84	77	80	-1.5	5.7	**2.3**	PC	-0.8	1.7	**5.6**	**0.07**	**1.43**
Grade 5 (n=24)	87	84	87	89	-3.7	-1.2	-2.7	PC	-3.8	-1.2	-2.6	0.61	
Fluency (WPM)	58	51	63	59	12.0	10.4	9.1	WI, FLU	11.8	8.1	9.5	0.72	

Table 2: Results of 1999-2000 8-month comparison of treatment groups' pretest scores and gains on each test, by grade; highlighted gains are significantly higher than in one or both other conditions

Subtest	Mean Pretest	Mean Pretest by Condition			Actual Gain			Significant Covariates	Gain Adjusted by Covariate(s)[1]			Main Effects p =	Effect Size
		Class	HT	RT	Class	HT	RT		Class	HT	RT		
Grades 2-3		(n = 39)	(n = 34)	(n = 58)					(n = 39)	(n = 34)	(n = 58)		
Word Attack (normed)	90	90	89	90	3.4	**7.3**	0.1	WA, WI	3.8	**6.8**	0.2	**0.02**	**0.55**
Grade 2 (n=36)	85	87	83	85	8.2	**11.0**	3.1	WA, WI	9.1	**10.4**	2.8	**0.07**	**0.61**
Grade 3 (n=37)	94	93	93	95	-1.1	**3.6**	-2.8	WA, WI	-1.1	**3.5**	-2.7	**0.10**	**0.78**
Word ID (normed)	90	90	90	90	0.9	1.4	0.6	WI, WC	1.0	0.9	0.8	0.97	
Grade 2 (n=36)	90	90	89	90	1.6	1.0	-0.7	WI, WC	1.6	0.4	-0.3	0.16	
Grade 3 (n=37)	91	90	91	90	0.3	1.8	1.8	WI, WC	0.4	1.6	1.8	0.53	
Word Comp (normed)	90	90	92	89	3.1	3.9	4.4	WI, WC	2.9	4.8	4.0	0.46	
Grade 2 (n=36)	89	90	90	88	5.6	4.4	4.4	WI, WC	5.7	5.1	4.0	0.73	
Grade 3 (n=37)	91	90	94	90	0.7	**3.4**	**4.3**	WI, WC	0.0	**4.6**	**3.9**	**0.02**	**0.56, 0.72**
Passage Comp (normed)	90	90	90	89	1.4	3.1	3.6	WC, PC	1.9	2.4	3.6	0.46	
Grade 2 (n=36)	90	91	89	89	1.9	2.0	2.3	WC, PC	2.5	1.4	2.2	0.90	
Grade 3 (n=37)	90	89	91	89	1.0	**4.1**	**4.8**	WC, PC	1.3	**3.4**	**5.0**	**0.14**	**0.55, 0.48**
Fluency (on grade level text)	27	28	29	25	29.3	34.4	27.6	WC, PC, FLU	29.8	33.7	27.8	0.20	
Grade 2 (n=36)	14	14	16	13	39.1	40.6	34.8	WC, PC, FLU	38.4	39.7	35.9	0.71	
Grade 3 (n=37)	40	42	44	38	19.9	28.1	20.4	WC, PC, FLU	20.9	28.1	19.8	0.15	

Table 3: Pretest and gain on each measure, by grade, classroom, and treatment group

			N	Word Attack		Word Identification		Word Comprehension		Passage Comprehension		Fluency WPM	
	Class	**Treatment**		**Pre**	**Gain**	**Pre**	**Gain**	**Pre**	**Gain**	**Pre**	**Gain**	**Pre**	**Gain**
Grade 2	**201**	control	2	96.0	17.0	97.5	-0.5	97.5	1.5	91.5	4.5	21.0	45.5
		Reading Tutor	9	82.3	2.6	87.6	0.7	87.3	3.7	88.1	2.2	12.7	43.2
	205	control	6	79.8	8.5	88.7	2.5	88.8	4.3	90.3	-3.3	12.2	34.7
		human (MB)	6	77.2	17.7	91.3	-0.5	91.7	9.7	91.2	5.2	20.5	49.8
	208	control	6	93.7	7.7	91.0	-0.5	88.7	7.8	91.2	6.0	12.0	42.7
		human (ME)	5	85.0	0.4	83.8	5.0	87.0	-4.0	85.0	2.0	8.8	32.4
	209	control	3	89.7	10.3	96.3	-1.0	95.0	7.0	92.0	3.0	16.7	45.3
		human (AC)	6	88.0	13.2	91.8	-0.8	92.0	6.2	91.5	-1.2	16.2	38.3
	211	Reading Tutor	10	86.6	5.0	93.8	-2.5	90.8	3.9	92.3	-1.3	17.6	27.5
	212	control	2	72.0	-3.5	75.5	11.0	79.5	5.0	88.5	1.0	9.5	26.0
		Reading Tutor	10	84.6	1.6	89.2	0.0	86.9	5.7	86.1	6.1	8.7	34.6
	Grade Total	control	19	86.6	8.2	90.2	1.6	89.7	5.6	90.8	1.9	13.5	39.1
		Reading Tutor	29	84.6	3.1	90.3	-0.7	88.4	4.4	88.9	2.3	13.0	34.8
		human	17	83.3	11.0	89.3	1.0	90.4	4.4	89.5	2.0	15.5	40.6
Grade 3	**301**	control	2	80.5	2.5	88.5	-0.5	90.0	1.0	89.0	2.5	35.0	17.0
		Reading Tutor	8	94.0	1.0	89.3	4.9	87.5	6.0	87.0	6.3	25.1	25.8
	303	control	2	93.0	-5.5	83.5	-2.0	79.5	9.5	83.5	-1.0	19.5	16.0
		Reading Tutor	10	94.7	-9.2	90.1	-1.2	89.0	2.9	87.8	5.9	42.6	12.1
	304	Reading Tutor	11	96.8	0.2	91.4	2.2	93.8	4.4	92.7	2.6	42.7	24.2
	305	control	5	93.0	-1.0	90.0	-1.0	85.8	7.6	87.8	2.4	33.6	29.0
		human (LN)	5	88.0	4.0	89.2	3.0	93.0	4.8	90.2	5.0	52.2	31.8
	309	control	5	93.4	3.2	88.2	1.8	89.4	-4.4	87.8	-2.0	41.2	19.6
		human (MM)	7	94.0	5.3	90.4	2.0	91.7	3.6	89.3	3.1	44.6	32.6
	310	control	6	98.0	-4.3	94.3	1.2	96.5	-4.0	93.3	2.5	58.0	14.8
		human (NJ)	5	97.4	1.0	91.6	2.4	96.6	2.2	91.8	5.2	33.4	18.2
	Grade Total	control	20	93.4	-1.0	90.1	0.3	89.7	0.6	89.2	1.0	41.6	19.9
		Reading Tutor	29	95.3	-2.8	90.3	1.8	90.4	4.3	89.4	4.8	37.8	20.4
		human	17	93.2	3.6	90.4	2.4	93.5	3.5	90.3	4.3	43.5	28.1

Table 4: Mean Session Times (minutes) and Mean Reading Rate (text words per minute)

	Second Grade		Third Grade	
	Reading Tutor	Human Tutor	Reading Tutor	Human Tutor
Total Work Time (minutes)	17.9	16.7	18.8	14.2
Waiting Time (minutes)	7.7	--	8.7	--
Assisted Reading Rate (wpm)				
Text Words / Elapsed Reading Time	14.4	20.5	20.4	52.1
Text Words / Net Reading Time (excluding waiting time)	28.1	--	40.8	--

Table 5: Reading Errors and Help Requests

	Second Grade			Third Grade		
Counts as text words per session:	Reading Tutor		Human Tutor	Reading Tutor		Human Tutor
Mean Reading Errors	13.8		17.4	17.6		22.1
Word-Level Help Requests	11.5		3.7	9.9		3.2
Sentence-Level Help Requests	6.5		0	0.8		0
Rates as percentage of text words:	Story chosen by Student	Tutor		Story chosen by Student	Tutor	
Reading Error Rate	0.02	0.07	0.09	0.03	0.05	0.05
Disposition of errors						
% Tutor corrected explicitly	55%	22%	81%	24%	18%	69%
% Tutor corrected incidentally	22%	18%	--	1%	7%	--
% Self corrected	7%	18%	12%	45%	17%	15%
% Student asks for help	11%	16%	--	0%	11%	--
% Uncorrected	6%	41%	7%	30%	47%	16%
Word-Level Help Request Rate	0.02	0.08	0.02	0.01	0.04	0.01
Sentence-Level Help Request Rate	0.04	0.03	--	0.003	0.008	--

Table 6: Categories of Tutor Intervention

	Second Grade			Third Grade		
	Reading Tutor		Human Tutor	Reading Tutor		Human Tutor
Who chose story:	Student	Tutor		Student	Tutor	
Overall Intervention Rate (interventions per word of text)	0.38	0.23	0.18	0.18	0.23	0.07
Total Reading Assistance						
% of all tutor interventions	83%	80%	57%	70%	78%	71%
Rate per text word	0.32	0.19	0.11	0.12	0.18	0.05
Pre-emptive						
% of all tutor interventions	33%	13%	--	34%	21%	--
Rate per text word	0.13	0.05	--	0.06	0.05	
Errors and Help						
% of all tutor interventions	39%	44%	57%	15%	43%	71%
Rate per text word	0.15	0.10	0.11	0.03	0.10	0.05
False Alarm						
% of all tutor interventions	10%	14%	--	20%	14%	--
Rate per text word	0.04	0.03	--	0.04	0.03	--
Praise/Backchanneling						
% of all tutor interventions	17%	20%	35%	30%	22%	22%
Rate per text word	0.07	0.05	0.06	0.05	0.05	0.02
Discussion of Meaning						
% of all tutor interventions	--	--	8%	--	--	7%
Rate per text word	--	--	0.01	--	--	0.00

Table 7: Categories of Reading Assistance

	Second Grade			Third Grade		
	Reading Tutor		Human Tutor	Reading Tutor		Human Tutor
	Self	Tutor		Self	Tutor	
Focus on Word	27%	28%	27%	23%	18%	22%
Read a Word	24%	18%	18%	29%	26%	46%
Read Whole Sentence	35%	30%	0%	24%	30%	0%
Exaggerated Sounding Out	9%	13%	14%	11%	17%	17%
Rhyme	1%	6%	1%	4%	3%	0%
Letter-Sound Correspondence	4%	7%	31%	8%	5%	11%
Letter-Sound Pattern Rule	--	--	7%	--	--	0%
Spell	--	--	1%	--	--	1%
Semantic Cue	--	--	1%	--	--	2%

Table 8: Comparison of process variables for Reading Tutor (RT) and human tutoring (HT), by grade

Process variable, data source (and how derived) [averaged by student to avoid bias; shown by grade and by RT room or HT initials]	**Grade 2**		**Grade 3**	
	Reading Tutor n=29	**Human tutor** n=17	**Reading Tutor** n=29	**Human tutor** n=17
Total number of sessions	**67 days**	**73 days**	**71 days**	**>> 61 days**
RT event database (days with any events) HT log (days with any logged activity)	90++RT20 1 54 RT211 56 RT212	67 AC 77 MB 77 ME	70 RT301 57--RT303 86 RT304	61 LN 62 MM 58 NJ
Story words seen per session	**122 words**	**<? 154 words**	**143 words**	**<< 262 words**
RT portfolio (#words of finished stories only!) HT log (#words in logged stories; prorated for never-finished stories based on # pages read)	120 RT201 108 RT211 135 RT212	112 AC 224 MB 120 ME	122-RT301 143 RT303 162 RT304	258 LN 313 MM 194-NJ
Level of stories finished, chosen (tutor/child)	**1.1(1.8/1.1)**	**<< 1.8**	**1.7(2.5/1.8)**	**<< 2.2**
RT portfolio (shows if finished and who chose; finished stories averaged a half level lower.) HT log (shows level, pages read, not who chose)	1.1 RT201 0.8 RT211 1.2 RT212	1.4 AC 2.8+MB 1.2 ME	1.4 RT301 2.0 RT303 1.7 RT304	2.3 LN 2.2 MM 2.2 NJ
Percentage of rereading	**30%**	**>> 19%**	**24%**	**>> 13%**
RT portfolio (% of finished stories read before) HT log (% of finished stories read before)	34% RT201 28% RT211 29% RT212	24% AC 11% MB 21% ME	25% RT301 18%-RT303 30% RT304	13% LN 18% MM 6% NJ
Percent of sessions with any writing activity	**38%**	**<< 64%**	**28%**	**<< 58%**
RT event logs (% of days with edit events) HT log (listed writing activities)	46% RT201 37% RT211 32% RT212	85% AC 37%--MB 70% ME	36% RT301 25% RT303 22% RT304	67% LN 60% MM 44%--NJ

Table 9: Partial correlations of gains with each other and with process variables, controlling for significant pretest covariates (**?**, *****, and ****** indicate respective significance levels of $p < .10$, $p < .05$, and $p < .01$)

		Word Attack Normed Score Gain	Word Identification Normed Gain	Word Comprehension Normed Gain	Passage Comprehension Normed Gain	Fluency Gain (WPM)
	Covariates:	WA, WI	WI, WC	WI, WC	WC, PC	WC,PC,FLU
GRADE 2:						
Word Attack	human tutors	1.000	0.059	0.428	-0.022	0.287
Normed Score Gain	Reading Tutor	1.000	0.314	-0.057	0.116	-0.152
Word Identification	human tutors	0.415	1.000	-0.091	-0.078	0.011
Normed Score Gain	Reading Tutor	0.193	1.000	**0.337?**	0.128	0.278
Word Comprehension	human tutors	0.394	-0.091	1.000	0.068	0.311
Normed Score Gain	Reading Tutor	-0.099	**0.337?**	1.000	**0.608****	0.055
Passage Comprehension	human tutors	-0.051	-0.039	0.049	1.000	**0.506?**
Normed Score Gain	Reading Tutor	0.033	**0.448***	**0.482***	1.000	0.010
Fluency Gain (WPM)	human tutors	0.271	0.096	0.254	**0.513***	1.000
	Reading Tutor	-0.060	**0.391***	0.032	0.006	1.000
Sessions	human tutors	-0.061	-0.260	0.244	0.196	0.349
	Reading Tutor	0.062	-0.025	-0.141	-0.131	0.297
Level	human tutors	0.375	0.248	0.378	**0.446?**	**0.523?**
	Reading Tutor	0.175	0.225	**0.328?**	**0.347?**	**0.497***
Re-reading	human tutors	-0.024	-0.317	**-0.448?**	-0.355	-0.405
	Reading Tutor	0.036	0.042	0.057	-0.274	-0.075
Writing	human tutors	-0.026	-0.140	-0.292	**-0.481?**	-0.417
	Reading Tutor	0.000	-0.226	0.009	-0.211	0.057
Words	human tutors	0.297	0.294	0.223	**0.652****	**0.577***
	Reading Tutor	0.035	**0.505****	**0.559****	0.312	0.324
GRADE 3:						
Word Attack	human tutors	1.000	0.210	-0.008	0.126	0.079
Normed Score Gain	Reading Tutor	1.000	**0.589****	0.292	0.196	**0.388***
Word Identification	human tutors	0.274	1.000	0.380	**0.518***	0.072
Normed Score Gain	Reading Tutor	**0.567****	1.000	0.075	0.254	**0.442***
Word Comprehension	human tutors	-0.049	0.380	1.000	**0.513***	0.315
Normed Score Gain	Reading Tutor	0.285	0.075	1.000	0.264	0.059
Passage Comprehension	human tutors	0.241	0.172	-0.073	1.000	-0.068
Normed Score Gain	Reading Tutor	0.028	0.234	0.098	1.000	**0.453***
Fluency Gain (WPM)	human tutors	-0.039	-0.097	0.264	-0.299	1.000
	Reading Tutor	**0.453***	**0.471***	0.200	**0.373?**	1.000
Sessions	human tutors	0.106	-0.341	0.205	-0.185	0.173
	Reading Tutor	**0.393***	0.222	0.188	-0.183	0.266
Level	human tutors	0.103	0.090	0.142	0.339	**0.657***
	Reading Tutor	0.002	0.117	-0.238	0.204	**0.337?**
Re-reading	human tutors	-0.056	0.152	0.045	0.095	0.173
	Reading Tutor	**0.510****	0.230	**0.433***	0.177	0.232
Writing	human tutors	0.238	0.092	0.228	0.139	**0.530?**
	Reading Tutor	-0.024	0.103	-0.113	-0.005	-0.156
Words	human tutors	0.106	0.418	0.168	0.388	**0.495?**
	Reading Tutor	0.107	0.031	-0.036	-0.032	0.192

Table 10: 1998 and 1999-2000 study summaries in National Reading Panel scheme (NRP, 2000)

	Spring 1998	1999-2000
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	12
Number of participants	72	144
Age	7-11	7-10
Grade	2, 4, 5	2, 3
Reading levels of participants	Beginning- Intermediate; WRMT normed pretest ~84, grade equivalent K to 5	Beginning- Intermediate; WRMT normed pretest ~90, grade equivalent K to 3
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 Number of participants lost per group during the study Was attrition greater for some groups that others?	72 started in larger study 5 moved 4 unavailable → 63 overall 24 using Reading Tutor	144 started 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 except human tutor pullout
Design of study	Random assignment matched by pretest within classroom	Random assignment matched by pretest within classroom, but no classroom had both Reading Tutor and human tutors
Describe all treatment and control conditions; be sure to describe nature and components of reading	1998 Reading Tutor; regular classroom instruction; commercial reading software	1999-2000 Reading Tutor; regular classroom instruction; individual tutoring by certified teachers

instruction provided to control group		
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. Human tutors used same texts. Reading Tutor inserted short factoids to introduce some new words.
Duration of treatments	Nominally 20-25 minutes per day, 5 days per week, for entire spring Actual usage ~13 minutes/session, 1 day in 4-8	Nominally 20 minutes per day, 5 days per week, for entire fall Actual usage close to nominal guidelines, but varied by room
Was fidelity in delivering treatment checked?	Weekly visits by Project LISTEN personnel	2-3x/week visits by Project LISTEN personnel, plus logs of tutor sessions
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	Grade	Grade Room (specific teacher/tutor)
List processes that were taught during training and measured during and at the end of training	Not applicable	Not applicable
List names of reading outcomes measured	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
List time points when dependent measures were assessed	January 1998 and May 1998	September 1999 and May 2000
Any reason to believe that treatment/control groups might not have been equivalent prior to treatments?	No; pretest scores matched well.	No; pretest scores matched well.
Were steps taken in statistical analyses to adjust for any lack of equivalence?	Yes; analysis of variance controlled for pretest scores.	
Result: normed score gains, adjusted by significant covariates	Passage Comprehension p=.106	Word Attack p = .017 Grade 3 Word Comprehension p=.018 Grade 3 Passage Comprehension p=.14
Difference: treatment mean minus control mean	PC: Reading Tutor > class by 4.3	WA: human tutors > computer by 6.6 Grade 3 WC: computer > class by 3.9, human tutors > class by 4.6 Grade 3 PC: computer > class by 3.7

Effect size	PC: .60	WA: .55 Grade 3 WC: .56 computer, .72 human Grade 3 PC: .48 computer, .34 human
Summary statistics used to derive effect size	PC Reading Tutor gains: 2.4 PC class gains: -1.9 PC average SD: 7.2	WA Reading Tutor gains: 0.2 WA human tutor gains: 6.8 WA average SD: 12.0
		Gr. 3 WC Reading Tutor gains: 3.9 Gr. 3 WC class gains: 0.0 Gr. 3 WC Reading Tutor & class SD: 6.9 Gr. 3 WC human tutor gains: 4.6 Gr. 3 WC human tutor & class SD: 6.4
		Gr. 3 PC Reading Tutor gains: 5.0 Gr. 3 PC class gains: 1.3 Gr. 3 PC Reading Tutor & class SD: 7.7 Gr. 3 PC human tutor gains: 3.4 Gr. 3 PC human tutor & class SD: 6.2
Number of people providing effect size information	Entire sample	Entire sample
Length of time to code study	Uncertain	Uncertain
Name of coder	Mostow, adapted from (G. Aist, 2000)	

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

Aist, G. (1998, December). Expanding A Time-Sensitive Conversational Architecture For Turn-Taking To Handle Content-Driven Interruption. *Proceedings of the International Conference on Speech and Language Processing (ICSLP98)*, Sydney, Australia, Paper 928.

---. (2000). *Helping Children Learn Vocabulary during Computer-Assisted Oral Reading*. Unpublished Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA.

---. (2001a). Factoids: Automatically constructing and administering vocabulary assistance and assessment. In J. D. Moore, C. L. Redfield & W. L. Johnson (Eds.), *Artificial Intelligence in Education: AI-ED in the Wired and Wireless Future* (pp. 234-245). San Antonio, Texas: Amsterdam: IOS Press.

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

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

---. (2002b, April 29). Helping Children Learn Vocabulary during Computer-Assisted Oral Reading: A Dissertation Summary [Poster presented as a Distinguished Finalist for the Outstanding Dissertation of the Year Award]. *47th Annual Convention of the International Reading Association*, San Francisco, CA.

Aist, G., & Mostow, J. (1997a, November). A time to be silent and a time to speak: Time-sensitive communicative actions in a reading tutor that listens. *AAAI Fall Symposium on Communicative Actions in Humans and Machines*, Boston, MA.

---. (1997b, October). When Speech Input is Not an Afterthought: A Reading Tutor that Listens. *Workshop on Perceptual User Interfaces*, Banff, Canada.

---. (1999, September). Measuring the Effects of Backchanneling in Computerized Oral Reading Tutoring. *Proceedings of the ESCA Workshop on Prosody and Dialog*, Eindhoven, Netherlands.

---. (2000, June). Improving story choice in a reading tutor that listens. *Proceedings of the Fifth International Conference on Intelligent Tutoring Systems (ITS'2000)*, Montreal, Canada, 645.

---. (in press). Faster, better task choice in a reading tutor that listens. In P. DelCloque & M. Holland (Eds.), *Speech Technology for Language Learning*. Lisse, The Netherlands: Swets & Zeitlinger Publishers.

Aist, G., Mostow, J., Tobin, B., Burkhead, P., Corbett, A., Cuneo, A., Junker, B., & Sklar, M. B. (2001). Computer-assisted oral reading helps third graders learn vocabulary better than a classroom control – about as well as one-on-one human-assisted oral reading. In J. D. Moore, C. L. Redfield & W. L. Johnson (Eds.), *Artificial Intelligence in Education: AI-ED in the Wired and Wireless Future* (pp. 267-277). San Antonio, Texas: Amsterdam: IOS Press.

Betts, E. A. (1946). *Foundations of Reading Instruction*. New York: American Book Company.

Clay, M. M. (1991). Why is an inservice programme for Reading Recovery teachers necessary? *Reading Horizons, 31*(5), 355-372.

Cunningham, A. E., & Stanovich, K. E. (1991). Tracking the Unique Effects of Print Exposure in Children: Associations with Vocabulary, General Knowledge, and Spelling. *Journal of Educational Psychology, 83*(2), 264-274.

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

Fielding, L., Wilson, P., & Anderson, R. (1986). A new focus on reading: The role of trade books in reading instruction. In T. Raphael & R. Reynolds (Eds.), *Contexts of Literacy*. New York: Longman.

Fogarty, J., Dabbish, L., Steck, D., & Mostow, J. (2001). Mining a database of reading mistakes: For what should an automated Reading Tutor listen? In J. D. Moore, C. L. Redfield & W. L. Johnson (Eds.), *Artificial Intelligence in Education: AI-ED in the Wired and Wireless Future* (pp. 422-433). San Antonio, Texas: Amsterdam: IOS Press.

Foorman, B. R., Francis, D. J., Fletcher, J. M., Mehta, P., & Schatschneider, C. (1998). The Role of Instruction in Learning To Read: Preventing Reading Failure in At-Risk Children. *Journal of Educational Psychology, 90*(1), 37-55.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Walz, L., & others. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review, 22*(1), 27-48.

Icabone, D., & Hannaford, A. (1986). Comparison of two methods of teaching unknown reading words to fourth graders: Microcomputer and tutor. *Educational Technology, 26*, 36-39.

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

Kamil, M. L., Mosenthal, P. B., Pearson, P. D., & Barr, R. (Eds.). (2000). *Handbook of Reading Research, Volume III*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Leinhardt, G., Zigmond, N., & Cooley, W. W. (1981). Reading instruction and its effects. *American Education Research Journal, 13*(3), 343-361.

Levy, B. A., Nicholls, A., & Kohen, D. (1993). Repeated readings: Process benefits for good and poor readers. *Journal of Experimental Child Psychology, 56*, 303-327.

Lundberg, I., & Olofsson, A. (1993). Can computer speech support reading comprehension? *Computers in Human Behaviour, 9*(2-3), 283-293.

McConkie, G. W. (1990). *Electronic Vocabulary Assistance Facilitates Reading Comprehension: Computer Aided Reading*.Unpublished manuscript.
Mostow, J., & Aist, G. (1999a, July). Authoring new material in a reading tutor that listens. *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99), Intelligent Systems Demonstration track*, Orlando, FL, 918-919.
---. (1999b). 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.
---. (1999c). Reading and Pronunciation Tutor (United States Patent No. 5,920,838), *US Patent and Trademark Office*.
---. (2001). Evaluating tutors that listen: An overview of Project LISTEN. In K. Forbus & P. Feltovich (Eds.), *Smart Machines in Education* (pp. 169-234). Menlo Park, CA: MIT/AAAI Press.
Mostow, J., Aist, G., Beck, J., Chalasani, R., Cuneo, A., Jia, P., & Kadaru, K. (2002, June 5-7). A La Recherche du Temps Perdu, or As Time Goes By: Where does the time go in a Reading Tutor that listens? *Proceedings of the Sixth International Conference on Intelligent Tutoring Systems (ITS'2002)*, Biarritz, France, 320-329.
Mostow, J., Aist, G., Bey, J., Burkhead, P., Cuneo, A., Rossbach, S., Tobin, B., Valeri, J., & Wilson, S. (2001). A hands-on demonstration of Project LISTEN's Reading Tutor and its embedded experiments (refereed demo). *Language Technologies 2001: The Second Meeting of the North American Chapter of the Association for Computational Linguistics.*, Pittsburgh, PA.
Mostow, J., Aist, G., Huang, C., Junker, B., Kennedy, R., Lan, H., IV, D. L., O'Connor, R., Tassone, R., Tobin, B., & Wierman, A. (in press). 4-Month Evaluation of a Learner-controlled Reading Tutor that Listens. In P. DelCloque & M. Holland (Eds.), *Speech Technology for Language Learning*. Lisse, The Netherlands: Swets & Zeitlinger Publishers.
Mostow, J., Beck, J., Winter, S. V., Wang, S., & Tobin, B. (2002, September 16-20). Predicting oral reading miscues. *Proceedings of the Seventh International Conference on Spoken Language Processing (ICSLP-02)*, Denver, CO.
Mostow, J., Hauptmann, A. G., Chase, L. L., & Roth, S. (1993, July). Towards a reading coach that listens: automated detection of oral reading errors. *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI93)*, Washington, DC, 392-397.
Mostow, J., Huang, C., & Tobin, B. (2001). Pause the Video: Quick but quantitative expert evaluation of tutorial choices in a Reading Tutor that listens. In J. D. Moore, C. L. Redfield & W. L. Johnson (Eds.), *Artificial Intelligence in Education: AI-ED in the Wired and Wireless Future* (pp. 343-353). San Antonio, Texas: Amsterdam: IOS Press.
Mostow, J., Roth, S. F., Hauptmann, A. G., & Kane, M. (1994, August). A prototype reading coach that listens [AAAI-94 Outstanding Paper Award]. *Proceedings of the Twelfth National Conference on Artificial Intelligence*, Seattle, WA, 785-792.
Nagy, W., Herman, P., & Anderson, R. (1985). Learning words from context. *Reading Research Quarterly, 20*, 233-253.
NCES. (2000). *National Assessment of Educational Progress*, from http://nces.ed.gov/nationsreportcard
NRP. (2000). *Report of the National Reading Panel: Teaching Children to Read*. Washington, DC: National Institute of Child Health & Human Development.
Olson, R., Foltz, G., & Wise, B. (1986). Reading instruction and remediation with the aid of computer speech. *Behavior Research Methods, Instruments, & Computers, 18*(2), 93-99.
Olson, R. K., & Wise, B. (1987). Computer speech in reading instruction. In D. Reinking (Ed.), *Computers and Reading: Issues for Theory and Practice*. New York: Teachers College Press.
Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing Reading Difficulties in Young Children*. Washington, D.C.: National Academy Press.
Wasik, B. A., & Slavin, R. E. (1993). Preventing early reading failure with one-to-one tutoring: A review of five programs. *Reading Research Quarterly, 28*(2), 178-200.
Wise, B. W. (1992). Whole Words and Decoding for Short-Term Learning: Comparisons on a "Talking-Computer" System. *Journal of Experimental Child Psychology, 54*(2), 147-167.
Wise, B. W., Olson, R., Anstett, M., Andrews, L., Terjak, M., Schneider, V., Kostuch, J., & Kriho, L. (1989). Implementing a long-term computerized remedial reading program with synthetic speech feedback: hardware, software and real-world issues. *Behavior Research Methods, Instruments, & Computers, 21*(2), 173-180.
Wise, B. W., & Olson, R. K. (1992). How Poor Readers and Spellers Use Interactive Speech in a Computerized Spelling Program. *Reading and Writing: An Interdisciplinary Journal, 4*(2), 145-163.
Wise, B. W., Ring, J., & Olson, R. K. (1999). Training phonological awareness with and without explicit attention to articulation. *Journal of Experimental Child Psychology, 72*(4), 271-304.
Woodcock, R. W. (1987). *Woodcock Reading Mastery Tests-Revised*. Circle Pines, MN: American Guidance Service.
---. (1998). *Woodcock Reading Mastery Tests - Revised (WRMT-R/NU)*. Circle Pines, Minnesota: American Guidance Service.