Computer-Guided Oral Reading versus Independent Practice:

Comparison of Sustained Silent Reading to an Automated Reading Tutor that Listens

RUNNING HEAD: Computer-Guided Oral Reading versus Independent Practice

Jack Mostow and Jessica Nelson-Taylor, Carnegie Mellon University

Joseph E. Beck, Worcester Polytechnic Institute

Abstract

A 7-month study of 178 students in grades 1-4 at two Blue Ribbon schools compared two daily 20-minute treatments. 88 students used the 2000-2001 version of Project LISTEN's Reading Tutor (www.cs.cmu.edu/~listen) in 10-computer labs, averaging 19 hours over the course of the year. The Reading Tutor served as a computerized implementation of the National Reading Panel's recommended guided oral reading instruction (NRP, 2000). The Reading Tutor listened to students read aloud, giving spoken and graphical help when it noticed them click for help, make a mistake, or get stuck. Students using the Reading Tutor averaged significantly higher gains across measures of reading ability, especially those involving word level skills (word identification, blending words, and spelling) than their matched classmates who spent that time doing Sustained Silent Reading (SSR) in their classrooms. Additionally, these students trended towards higher gains in fluency and reading comprehension. Overall, use of the Reading Tutor resulted in the types of improvement that would be expected from guided oral reading, but with the benefit of scalability, a problem for human-guided oral reading practice.

Keywords: guided oral reading, fluency, Reading Tutor, computer-assisted reading, Sustained Silent Reading

Corresponding author:

Jack Mostow, CMU-RI-NSH 4103, 5000 Forbes Avenue, Pittsburgh, PA 15213-3890, USA

Telephone 412-268-1330, FAX 412-268-6436, email mostow@cs.cmu.edu

This paper presents an empirical evaluation of computer-guided oral reading. Given the evidence-based recommendation of the National Reading panel to implement guided oral reading as an effective educational strategy to improve reading and fluency (NRP, 2000), we address the question of whether *computer*-guided oral reading is also an effective means of improving reading and fluency. Computer-guided oral reading has the potential advantage of being both a scalable and well-controlled form of reading instruction that allows students more time to practice reading aloud than they would in a traditional classroom environment.

Guided oral reading

Guided oral reading consists of reading aloud with individual attention and feedback, such as selecting appropriate text to read, helping decode hard words, correcting reading mistakes, providing encouragement, and prompting students to reread text to improve their fluency. The National Reading Panel found human-guided oral reading effective across a number of reading measures. "Guided oral reading procedures such as repeated reading … tended to improve word recognition, fluency (speed and accuracy of oral reading), and comprehension with most groups" (NRP, 2000, p. 3.38).

The National Reading Panel report does not explicitly define "guided oral reading," but mentions it mostly (though not exclusively) in the context of repeated reading, in which a

student reads the same text several times to build fluency on it. However, a subsequent review of fluency instruction (Kuhn & Stahl, 2003) found both rereading and wide reading effective, without a clear advantage of either over the other.

Guided oral reading confers benefits, but human guidance incurs costs. Peers may lack patience or skill. Parents may lack time or ability. Teachers are too busy teaching to spend very much time individual guiding students' oral reading. Even assuming a class size of only 20 students and a 2-hour daily block of Language Arts time, by devoting the entire block to individually guided reading the teacher would still listen to each student read only 6 minutes per day. More realistically, one study of teachers' classroom assistance for individual oral reading (Campbell, 1988) found that children received an average of less than five minutes per week of such assistance.

What about automated guidance? Can a computer program with speech recognition capabilities guide oral reading effectively enough to improve reading skills? If so, it would be an economical, scalable way to provide individually guided oral reading, free of human limitations on availability, patience, and skill (though of course limited by what computers can do). Thus computer-guided oral reading merits study.

The Reading Tutor: Computer-guided oral reading

This study evaluates a program that uses speech recognition to listen to children read aloud (Mostow & Aist, 1997; Mostow et al., 1994) – namely, the 2000-2001 version of Project LISTEN's Reading Tutor (Mostow & Aist, 2001). It is important to acknowledge that its speech recognition was far from perfect. It was over 95% accurate in accepting correctly read words, but mediocre in detecting miscues. Thus it behaved somewhat like an infinitely

patient but somewhat hard-of-hearing grandparent who ignores most miscues, but can read fluently and expressively and help when a child gets stuck or encounters a difficult word.

A session with the Reading Tutor consisted of logging in and reading one or more stories aloud, sometimes preceded by a preview activity or followed by a review, such as an introduction or a test of new vocabulary. Story text came from various sources, including *Weekly Reader* (a newspaper distributed to children in many American schools, with grade-levelled editions) minus pictures, Aesop's Fables and other public-domain materials available on-line from Project Guttenberg, and stories developed by project members, with story level and genre assigned manually. The Reading Tutor displayed each story incrementally, adding one sentence at a time, and listened to the student read it aloud, providing spoken and graphical assistance when it noticed the student click for help, hesitate, get stuck, skip a word, misread a word, or encounter a word likely to be misread (Mostow & Aist, 1999b).

The key features of the Reading Tutor were (a) designed for automaticity and consequently scalability and (b) based on empirically supported principles of instruction.

Automation/Scalability Features:

- *Automated interactive tutorials*. Students enrolled on the Reading Tutor at the start of the year, and received automated interactive tutorials on how to operate the Reading Tutor, how to use the keyboard, and (once they reached a third grade level in the Reading Tutor) how to author stories by typing, editing, and narrating them.
- *Automated student skill levelling*. Story levels ranged from kindergarten (K) to seventh grade (G). The original levels, corresponding to grades 1, 2, and 3, were

named A, B, and C to disguise their grade-equivalent independent reading levels so as to avoid stigmatizing below-grade-level readers; levels K (kindergarten) and D-G (grades 4-7) were added later. Based on age, the Reading Tutor placed the student at an initial reading level, which it then adjusted up (or down) based on the student's assisted reading rate, measured as the number of text words per minute accepted as read correctly (Aist, 2002).

- *Simple scripting language for adding content*. Stories, previews, and reviews were expressed in a simple scripting language as sequences of a few types of student activities: reading, listening, spelling, picking, editing, and narrating. This "activity language" was used to express reading, writing, and practice activities, interactive tutorials, student opinion polls, and embedded "invisible experiments" to evaluate various interventions (Mostow, 2008; Mostow et al., 2001). The Appendix describes the step types and interventions.
- *Constrained randomization*. Tutor choices of stories, help types, or other activities were first constrained by feasibility and requisite conditions (e.g. correct reading level, feasible help type for the given word) and then chosen randomly from the set of appropriate options. This simple mechanism generated diverse tutor behaviors.

Instructional Principles:

The design of the 2000-2001 version of the Reading Tutor is detailed elsewhere (Mostow & Aist, 1999b, 2001), but its underlying instructional principles can be summarized as follows:

- *Shared control*. Control was shared between the Reading Tutor and student in order to balance the motivational advantages of learner control (Lepper et al., 1993)

against its disadvantages. Examples of decisions at three different levels illustrate this principle. At the level of story choice, the Reading Tutor took turns with the student at picking stories to read, so as to prevent poor readers from merely reading the same easy stories over and over (Aist & Mostow, 2008). At the level of previews and reviews, the Trouble Word Review focused on words on which it thought the student needed help, but let the student choose among them. At the level of help on an individual word or sentence, the Reading Tutor gave help both when it thought the student needed help, and when the student asked for help. As one student said, "I like reading with the computer because when you don't know a word, you can click on it. You can't click on your teacher." This remark nicely captures not only the value of such learner control but the informal nature of the student's relationship to the Reading Tutor as opposed to an adult authority figure.

- *Wide reading and re-reading*. The Reading Tutor and student took turns picking stories. On its turn, the Reading Tutor always picked a new story (i.e. that the student had never completed), but the student could choose to reread an old story. Thus all story rereading was strictly voluntary. In contrast, guided oral reading reported in the literature often involves repeated reading, where a student reads the same text several times to build fluency on it. Although the Reading Tutor let students reread individual sentences, or choose to reread a story they had read before, it did not prompt the student to reread the same text several times to build fluency.

- *Phonological instruction.* Typical as-needed assistance included reading a word or sentence aloud, sounding it out, giving a "rhymes with" or "starts like" analogy, or pronouncing an individual grapheme ("P H here makes the sound /F/"). Following such assistance, the student could then reread part or all of the sentence, or click "Go" to go on. Mostow et al. (1994) showed that a more rudimentary form of such assistance enabled students to read and comprehend stories above their independent reading level.

Study Design

The study reported here investigates the effects of computer-guided oral reading in students learning to read in their native language (English) on a variety of reading measures. The National Reading Panel proposed "a new research convention for methodological studies with students in the second grade or higher. The amount of gain attributable to reading alone should be the baseline comparison against which the efficacy of instructional procedures is tested. If an instructional method does better than reading alone, it would be safe to conclude that method works" (NRP, 2000, Ch. 3, p. 27). To evaluate computer-guided oral reading, as instantiated in the 2000-2001 version of the Reading Tutor, we compared it to independent practice as implemented in Sustained Silent Reading (SSR). Sustained silent reading is a current practice in many schools, including the schools in this experiment. We compared reading gains across a number of tests in two comparably skilled groups of students in grades 1-4 who either engaged in the usual SSR activity (control group) or who instead used the Reading Tutor (experimental group) during the block time normally allotted for SSR.

Methods

Participants

The participants initially selected for the study were 193 students in grades 1-4 at two Blue Ribbon National Schools of Excellence (www.ed.gov, 2002) in the same school district. The district was suburban and affluent, with 90% of the student body white, 8% black, fewer than 1% Hispanic, and only 10% eligible for free or reduced lunch. 178 students (92%) completed the study: 75 in grade 1, 39 in grade 2, 35 in grade 3, and 29 in grade 4. Of these 178 students, 55 were special needs students (identified by teacher questionnaires as receiving individual services). Special needs are by definition individual to each student, and our data do not specify their nature. Compared to other students, the special needs students read more poorly and learned more slowly.

We did not prescribe selection criteria for the schools' reading specialists to use in deciding which students would participate in the study. Consequently, one school's entire first grade class participated, whereas in the other grades, the reading specialists selected the poorer readers from each grade, excluding the very lowest readers who met with the reading specialists during the time allotted for silent reading. To ensure that the comparison of instruction types was not affected by the selection, students were matched for skill and classroom between treatment conditions.

Assignment to treatment

A computerized matching program paired each student with another student in the same classroom, based on a pre-test Total Reading Composite score from the Woodcock Reading Mastery Test, Form G. Two non-readers, defined by raw scores below 2 on Word Attack and below 4 on Word Identification, were paired with each other in the same classroom.

Finally, one student in each pair was randomly assigned to the Sustained Silent Reading (SSR) condition and the other was assigned to the Reading Tutor (RT) condition. In classes with odd numbers of students, the median-ranked student was left unpaired, so as to keep the treatment groups well-matched. Pairing was done to improve the similarity of the two treatment groups, but paired students were not treated as paired samples in statistical analyses. Of the 178 students who completed the study, 90 were assigned to the SSR condition, and 88 used the Reading Tutor.

Apparatus

It is customary in reporting educational technology studies to describe the equipment used. The 2000-2001 Reading Tutors at the two labs ran under Windows NT™ on ordinary personal computers with 733 megahertz Pentium III™ processors, 128 megabytes of memory, 20-gigabyte disks, compact disk readers used to install software updates, 17" or 19" monitors operating at 1024 x 768 resolution, full-duplex sound cards, and connections to the school local area network. Students wore headsets with earphones and noise-cancelling close-talking microphones to speak and listen to the Reading Tutor. Automated scripts ran nightly on each computer to transmit selected data to the Project LISTEN lab over the Internet. At the end of the school year, project staff harvested all the Reading Tutor data by hand.

Treatments

The study compared two treatments – Sustained Silent Reading, and the Reading Tutor. Additional reading instruction during the rest of the school day was identical for both treatment groups.

Sustained Silent Reading (SSR). SSR (Collins, 1980) took place in the classroom, and had already been implemented in prior years at both schools. "Students should read silently every day, choose their own books, have uninterrupted time to read, be able to choose not to finish a book, observe the teacher modeling good reading habits, and not be required to take tests or write book reports on what they read" (Gardiner, 2001). SSR nominally consisted of independent reading of student-chosen books during a daily 20-25 minute period reserved for that purpose. Thus it is reasonable to infer that these books were at students' independent reading level.

What about students not yet able or willing to read independently? In some first and second grade classrooms, SSR included teacher read-aloud. Teacher read-aloud is a predictable and relatively well-specified adaptation for students not yet ready to read.

Thus we studied SSR as actually implemented, including occasional deviations from the nominal standard. This condition has both ecological and practical validity.

Reading Tutor (RT). During SSR time, students assigned to use the Reading Tutor did so in a 10-computer lab at each school. A school-provided monitor escorted students to and from the lab and supervised them there. As mentioned earlier, an automated interactive tutorial in the Reading Tutor itself trained students on how to operate it, the Reading Tutor took turns with students picking what to read next, and it chose texts for each student based on its estimate of that student's reading level, which it adjusted automatically based on assisted reading rate as a rough indicator of oral reading fluency.

Treatment Fidelity

We used questionnaires to ask teachers and lab monitors what students did during SSR time, both in class and in the Reading Tutor lab. To monitor treatment fidelity, we

complemented these instruments with in-school observations of both treatment conditions by the Educational Research Field Coordinator. These observations confirmed the limited reliability of questionnaires. In one case, they identified a class where students assigned to the SSR condition were attending a computer lab (not the Reading Tutor lab) once a week, and where the teacher was reading one-on-one with students during SSR time. We clarified the study design to school personnel, and believe that treatment implementation generally conformed to the study design.

The Reading Tutor collected several additional types of data for students who used it. This data included digital recordings of students' oral reading, the time-aligned output of the speech recognizer, the latencies between successive words of text (Beck et al., 2004; Mostow & Aist, 1997), the text of stories that students authored in the Reading Tutor, detailed logs of their interactions with the Reading Tutor, thousands of ratings (of interest, length, and difficulty) for stories the student had just finished reading, and an end-of-year questionnaire on students' attitudes toward the Reading Tutor, reading in general, and different types of stories. We subsequently parsed the detailed log data into a database (Mostow, Beck, et al., 2002) to facilitate the analysis of such questions as how student time was distributed (Mostow, Aist, et al., 2002) and when they requested help (Beck et al., 2003).

Table 1 summarizes the time allocation of all 88 students who used the Reading Tutor. It shows total time and the categories where most of it was spent – reading, previews and reviews, and writing. These categories are indicative rather than exhaustive, so they do not add up to the total. Table 1 also shows the percentage of time that students spent picking stories. This percentage turned out to be an interesting indicator because it was

controlled by student behavior. Time allocation was similar for boys and girls. Total time on the Reading Tutor ranged from 10.5 hours to 27 hours. Time spent reading (that is, in Read steps) ranged from 5.5 hours to 18.3 hours.

[[INSERT NEAR HERE:

Table 1: *Time allocation of all 88 students who used the Reading Tutor*]]

Experimental Measures

Pre-test and post-test scores were collected for the following tests:

- Comprehensive Test of Phonological Processing (CTOPP) (Wagner et al., 1999). The Blending Words (n=174), Elision, and Rapid Letter Naming subtests of the CTOPP were administered to measure phonemic and letter awareness. We analyzed the raw scores.
- Woodcock Reading Mastery Test (WRMT) (Woodcock, 1998). The Word Identification, Word Attack, Word Comprehension, and Passage Comprehension subtests of the WRMT, Form G, were administered to measure both lower-level and higher-level reading processes. Scores from the Word Identification and Word Attack subtests form a "Basic Skills" cluster score, and scores from the two comprehension measures form a "Reading Comprehension" cluster score. All scores combined form the Total Reading Composite score (TRC), the score on which students in the two treatments were matched at pre-test. For the WRMT tests and subtests, our analyses used W scores, which are "useful for marking an individual's progress over time" and comprise an equal-interval scale across grades (Jaffe, 2009).

- Oral Reading Fluency. Oral reading fluency, measured in words read correctly per minute, was recorded for three passages at the grade level of the student. We used the mean rate on the three passages as the fluency score.
- Spelling. To measure spelling accuracy, we used the Test of Written Spelling (Larsen et al., 1999).
- Reading Attitude. To measure students' attitudes toward recreational and academic reading, we used the Elementary Reading Attitude Survey (ERAS) (McKenna et al., 1995).

Pre-test scores on all tests were collected in October. Post-test scores for the WRMT and oral reading fluency were collected in April, and all other post-test scores were collected in May. School personnel, augmented by substitute teachers we hired, administered the CTOPP, WRMT, and fluency tests individually. The spelling test and ERAS were administered by teachers in class or by testers in small groups. The Educational Research Field Coordinator trained and supervised the testers.

Results

We now discuss the results detailed in Table 2 and Table 3. For each test, Table 2 shows the SSR and Reading Tutor groups' mean and standard deviation for pretest, posttest, and gain, while Table 3 shows their covariate-adjusted gains, the difference between them, two measures of effect size for this difference, and its statistical significance.

[[INSERT NEAR HERE: Table 2: *Summary of Raw Gains*]]

[[INSERT NEAR HERE: Table 3: *Summary of Treatment Effects*]]

Pre-test

The two treatment groups were similar in ability, grade, and special needs status. There were no statistically significant differences in any pre-test scores between the two groups – for all tests, $F(1,176) < .35$, $p > .55$. There was no significant difference between the mean grade level for the RT group ($M = 2.07$, $SD = 1.10$) and the SSR group ($M = 2.14$, $SD = 1.16$), $t(176) = .412$, $p = .68$. There was also no significant difference in the percentage of special needs students between the RT group (35.3%) and the SSR group (26.7%), $X^2(1, N=178) = 1.5$, $p > .26$.

There was, however, a higher percentage of males in the Reading Tutor group (58.0%) than in the SSR group (42.2%), $X^2(1, N=178) = 4.4$, $p = .05$. In addition, the distribution of student gender across grade levels trended toward being different for the two treatment groups, , $F(1,176) = 3.8$, $p = .08$. In the Reading Tutor group, the mean grade of the males ($M = 2.25$, $SD = 1.19$) was higher than the females ($M = 1.96$, $SD = 1.04$), whereas the grade levels of males ($M = 2.16$, $SD = 1.19$) and females ($M = 2.11$, $SD = 1.13$) in the SSR group were closer to being equal.

Despite this trend toward an interaction between gender and treatment condition for grade level, this difference did not translate to similar interactions for any of the pre-test scores – the closest pre-test to showing a gender x treatment interaction was the spelling test, $F(1,172) = 2.22$, $p = .14$. For all other pre-tests $F(1, 172) < .92$, $p > .34$. In addition, only one pre-test showed a significant effect of gender: the Elementary Reading Attitudes Survey (ERAS), $F(1, 176) = 12.26$, $p = .001$. Females had higher scores (more positive attitudes toward reading) ($M = 64.66$, $SD = 9.86$) than males ($M = 58.56$, $SD = 11.75$).

Gains

Except for oral reading fluency, pre-test scores for every test were significantly negatively correlated ($p < .01$) with gain scores, i.e., students with lower pre-test scores tended to show higher gains. Therefore we used an ANCOVA with pre-test scores as a covariate to evaluate gain scores for every test except oral reading fluency, for which we conducted a simple ANOVA on gain scores instead of an ANCOVA. We considered using scores on the Elementary Reading Attitudes Survey as a covariate, but they were strikingly uncorrelated with any of the gain scores.

Treatment, gender, and the treatment by gender interaction were fixed factors. Including gender and the interaction between gender and treatment as fixed factors allowed us to identify any differences between treatment groups that may stem from greater gains in only males or only females and may not be a direct result of treatment, a concern because of the difference in gender distributions over the two treatment groups. However, the interpretation of any such effects as being due solely to gender is limited by the fact that there may be a relationship between gender and grade level in the RT group.

We computed partial Eta-squared for the treatment conditions as a measure of effect size. Partial-eta squared describes the proportion of variance explained by the treatment *after* the other factors are removed from the total non-error variation. We go by the rule of thumb that .0099 is a small effect, .0588 is a medium effect, and .1379 is a large effect. We additionally report Cohen's *d* for significant treatment effects, computed as the difference in adjusted means (estimated marginal means) normalized by the root mean squared error of the ANCOVA. We use the rule of thumb that .2 is a small effect size, .5 is medium, and .8

is large (Cohen, 1988). Slight variability in the number of students completing each test is due to absences, teachers not administering tests, and/or other reasons.

CTOPP: Blending Words (Raw Scores)

Students using the Reading Tutor outgained students in the SSR condition, $F(1, 169) = 5.02$, $p < .05$, partial $\eta^2 = .029$, $d = .34$. Estimated marginal mean gains for the RT group were $M = 4.04$ ($SE = .31$) and for the SSR group were $M = 3.05$ ($SE = .31$). There was no significant effect of gender, $F(1, 169) = 1.01$, $p = .32$, and no interaction between gender and treatment $F(1, 169) = .29$, $p = .85$. Overall, this result can be considered a small but reliable benefit for CTOPP blending scores for the RT group.

CTOPP: Elision (Raw Scores)

There were no statistically significant effects on gains in the CTOPP elision test for treatment, $F(1, 169) = 2.49$, $p = .12$, or gender, $F(1, 169) = 1.04$, $p = .31$, nor was there a treatment x gender interaction, $F(1, 169) = .00$, $p = .95$.

CTOPP: Rapid Letter Naming (Raw Scores)

There were no statistically significant effects on gains in the CTOPP rapid letter naming test for treatment, $F(1, 168) = 2.20$, $p = .14$, or gender, $F(1, 168) = 2.12$, $p = .15$, nor was there a treatment x gender interaction, $F(1, 168) = .20$, $p = .66$.

WRMT: Word Identification (W Scores)

Students using the Reading Tutor outgained students in the SSR condition, $F(1, 173) = 90.75$, $p < .001$, partial $\eta^2 = .344$, $d = 1.45$. Estimated marginal mean gains for the RT group were $M = 39.53$ ($SE = 2.18$) and for the SSR group were $M = 10.29$ ($SE = 2.16$). There was no significant effect of gender, $F(1, 173) = 1.32$, $p = .25$, but there was a trend towards an interaction between gender and treatment $F(1, 173) = .3.57$, $p = .06$. Estimated

marginal means showed that boys in the reading tutor group outgained the girls (M = 44.20, SE = 3.32 vs. M = 34.86, SE = 2.83), whereas in the SSR group, the girls slightly outgained the boys (M = 11.43, SE = 2.81 vs. M = 9.15, SE = 3.29). Overall, the benefit of the Reading Tutor on Word Identification was quite large for both boys and girls.

WRMT: Word Attack (W Scores)

There were no statistically significant effects on gains in the WRMT Word Attack test for treatment, $F(1, 173) = .18$, $p = .67$, or gender, $F(1, 173) = .09$, $p = .77$, nor was there a treatment x gender interaction, $F(1, 173) = .02$, $p = .88$.

WRMT: Basic Skills Cluster (W Scores)

Students using the Reading Tutor outgained students in the SSR condition, $F(1, 173) = 65.28$, $p < .001$, partial $\eta^2 = .274$, $d = 1.23$. Estimated marginal mean gains for the RT group were M = 27.93 (SE = 1.24) and for the SSR group were M = 13.80 (SE = 1.23). There was no significant effect of gender, $F(1, 173) = .89$, $p = .35$ but there was a trend towards an interaction between gender and treatment $F(1, 173) = 2.88$, $p = .09$. Since the Basic Skills subtest is simply a combination of the Word Identification and Word Attack subtests, the results mirror the large effect seen in the Word Identification subtest but are slightly smaller (however, the effect size is still very large for this combined score).

WRMT: Word Comprehension (W Scores)

There were no statistically significant effects on gains in the WRMT Word Comprehension test for treatment, $F(1, 173) = 2.49$, $p = .17$, or gender, $F(1, 173) = .05$, $p = .83$, but there was a trend towards an interaction between gender and treatment, $F(1, 173) = 3.47$, $p = .06$, partial $\eta^2 = .020$. Males slightly outgained females in the Reading Tutor group, with estimated marginal means of 18.61 (SE = 1.00) compared to 16.84 (SE = 1.17). In the SSR

group, the reverse was true with an estimated marginal mean of 17.15 (*SE* = .99) for the females and 14.90 (*SE* = 1.16) for the males. This interaction follows the same general pattern as for the Word Identification subtest.

WRMT: Passage Comprehension (W Scores)

There were no statistically significant effects on gains in the WRMT Passage Comprehension test for treatment, $F(1, 173) = 2.19$, $p = .14$, or gender, $F(1, 173) = 1.50$, $p = .22$, nor was there a treatment x gender interaction, $F(1, 173 = .21$, $p = .65$).

WRMT: Reading Comprehension Cluster (W Scores)

There was a trend towards a significant, but small, effect of treatment in the Reading Comprehension cluster (a combination of Word and Passage Comprehension scores) in which students using the Reading Tutor outgained the students in the SSR condition, $F(1, 173) = 3.34$, $p = .07$, partial $\eta^2 = .020$, $d = .28$. Estimated marginal means show a gain in the RT group of 18.21 (*SE* = .74) compared to 16.31 (*SE* = .73) for the SSR group. There was no significant effect of gender, $F(1, 173) = 1.13$, $p = .29$, and no interaction between gender and treatment, $F(1, 173) = 1.91$, $p = .17$.

WRMT: Total Reading Composite Score (W Scores)

The Total Reading Composite Score comprises all of the subtests of the Woodcock Reading Mastery Test reported above. Students using the Reading Tutor considerably outgained students in the SSR condition with estimated marginal mean gains of 23.09 (*SE* = 78) for the RT group compared to 15.04 (*SE* = .77) for the SSR group, $F(1, 173) = 53.62$, $p < .001$, partial $\eta^2 = .237$, $d = 1.11$. There was no significant effect of gender, $F(1, 173) = .078$, $p = .78$, but there was a nearly significant (but small) interaction between gender and treatment, $F(1, 173) = 3.86$, $p = .051$, partial $\eta^2 = .022$. Again, since this score is a composite of all of

the subtests, the pattern of the interaction is similar to the pattern for the Word Identification test: estimated marginal means showed that boys in the reading tutor group outgained the girls (M = 24.33, SE = 1.02 vs. M = 21.86, SE = 1.19), whereas in the SSR group, the girls slightly outgained the boys (M = 15.97, SE = 1.01 vs. M = 14.11, SE = 1.18).

Oral Fluency (Mean words read correctly per minute)

Because there was not a significant correlation between pre-test fluency measures and oral fluency gains, we evaluated oral fluency gains via an ANOVA of gain scores, with treatment, gender, and the interaction between treatment and gender as the fixed effects. There was a trend towards a significant effect of treatment in which students using the Reading Tutor outgained the SSR group, $F(1, 171) = 2.91$, $p = .09$, partial partial $\eta^2 = .017$, $d = .26$. Estimated marginal means for the RT group gains were 30.29 additional words per minute (SE = 1.94) compared to a gain of 25.61 words per minute for the SSR group (SE = 1.94). There was no significant effect of gender, $F(1, 171) = 2.76$, $p = .10$ and no interaction between gender and condition, $F(1, 171) = .75$, $p = .39$.

Test of Written Spelling

Students using the Reading Tutor outgained students' spelling in the SSR condition, $F(1, 158) = 4.93$, $p < .05$, partial $\eta^2 = .03$, $d = .35$. Estimated marginal mean gains for the RT group were M = 6.36 (SE = .36) and for the SSR group were M = 5.24 (SE = .36). There was no significant effect of gender, $F(1, 158) = .34$, $p = .56$, and no treatment x gender interaction, $F(1, 158) = 2.27$, $p = .13$.

Elementary Reading Attitude Survey

Students overall decreased in their scores on the ERAS from pre-test to post-test across both groups (mean decrease of 2.87) – a sadly typical finding (McKenna et al., 1995). There

was no significant effect of treatment $F(1, 160) = .00$, $p = .98$, or gender $F(1, 160) = 2.50$, $p = .12$, and there was no interaction between treatment and gender, $F(1, 160) = .41$, $p = .52$.

Relation to Other Work

Prior to this study, controlled evaluations of the 1998 and 1999 versions of the Reading Tutor compared gains to those achieved by other treatments. In both studies, students took turns using a single Reading Tutor computer in their classroom. 17 children in grade 2, 4, and 5 who used the 1998 Reading Tutor for a few hours over the course of four months achieved significantly ($p < 0.002$) higher gains in Passage Comprehension on the Woodcock Reading Mastery Test-Revised (Woodcock, 1998) than their statistically matched classmates who spent that time in regular classroom instruction (Mostow, Aist, et al., 2008). In the 1999-2000 study (Mostow et al., 2003), a human-tutored group significantly outgained the Reading Tutor group only in Word Attack. 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), with a similar trend in Passage Comprehension ($p=.14$, respective effect sizes .48 and .34).

In both prior studies, the schedule of tutoring rotated so as to spare students from consistently missing the same subject. Consequently the classroom instruction replaced by tutoring included a combination of reading and non-reading instruction in unknown proportions that varied by student and did not, therefore, provide an equal-time comparison to classroom reading instruction. Moreover, neither classroom reading instruction nor human tutoring is a replicable comparison treatment; they vary from school to school, from classroom to classroom, and even from day to day, for example

whenever there is a substitute teacher. In contrast, the current study compares the Reading Tutor to a better-defined control treatment.

Subsequently, various third-party studies have published peer-reviewed controlled evaluations of the Reading Tutor compared to various alternatives, but all were with English language learners. They showed, in some cases, comparable gains in reading measures to control conditions (Cunningham, 2006; Korsah et al., 2010; Reeder et al., 2009), and, in other cases, greater gains than the control conditions – specifically in fluency (Korsah et al., 2010; Poulsen et al., 2007; Weber & Bali, 2010) and timed sight word recognition (Poulsen et al., 2007). Reported effect sizes ranged from .55 to 1.27 in these studies. The present study differs in that it examines the effects of Reading Tutor use by native English speakers at two Blue Ribbon Schools of Excellence, who were presumably already receiving high-quality reading instruction.

A few other research groups have developed speech recognition based programs that purport to help children learn to read (Adams, 2006; Hagen et al., 2007; Kantor et al., 2012; Williams et al., 2008; Wise et al., 2008). However, peer-reviewed, controlled evaluations of their effects on reading gains are rare. A 17-week quasi-experimental study of 410 mainstream students in grades 2-5 at two schools found significantly greater fluency gains (ES ranging from .53 in grade 2 to 0.26 in grade 5) by those who read with a speech recognition based reading tutor than those who did not (Adams, 2011). The present study rigorously compared the Reading Tutor to a well-defined, widely used alternative treatment, not merely a no-extra-treatment control condition.

Finally, Campuzano et al. (2009) reported on a large-scale evaluation of four commercial reading software products for grade 1 and two for 4, using as respective outcome measures

the version-9 (SESAT, 1996) and version-10 (SAT-10, 2003) reading batteries of the Stanford Achievement Test, as well as district-administered standardized tests such as the Iowa Tests of Basic Skills. They found no overall significant effect; one fourth grade product made a 2-point difference in NCE scores. In comparison, the study reported here used more sensitive outcome measures and a much smaller sample of students.

Discussion

Two groups of students matched on their Total Reading Composite scores from the Woodcock Reading Mastery Test scores spent 20-25 minutes a day for seven months either engaging in sustained silent reading (SSR group) or using a computer reading tutor (RT group). These students were pre-tested and post-tested on a variety of reading measures. In no case did the SSR group outperform the RT group. Students using the Reading Tutor, however, showed greater gains in reading skills across a variety of measures. The most notable effect was a huge gain in Word Identification skills (the ability to correctly read real words) in the Reading Tutor group compared to the SSR group (Cohen's $d = 1.45$). This effect carried over into the WRMT's composite Basic Skills score ($d = 1.23$) and to the Total Reading Composite score ($d = 1.11$), both of which incorporate the Word Identification score.

An increased ability to correctly read words could stem from improvement in phonemic awareness or phonological processing, and we see some evidence from our measures that *lexically-driven* phonological processing contributes to improvement in our sample, but little evidence that more flexible, basic phonological processing is the basis for improvement. Our tests of phonological processing included the Blending Words, Elision, and Rapid Letter Naming subtests of the CTOPP. Of these measures, students using the Reading Tutor

outgained the SSR group only on the Blending Words subtest (d = .34). In addition, there was no significant difference between the two groups in Word Attack gains (the ability to read non-words or very low frequency words), which would also be expected to result from an improvement in basic phonological processing. The Blending Words subtest measures a student's ability to combine sounds to form real words (e.g. the student must combine the orally produced sounds *can-* and *-dy* to make *candy*). This is similar to the type of help a student would receive from the Reading Tutor if he or she were struggling to read a word. Therefore, the Reading Tutor group was given specific practice in this way of breaking down words into ordered phonological chunks, and this is likely to have improved their relative gains on the Blending Words test, and to have improved their ability to correctly read real words.

We also see significant improvement in spelling ability in the Reading Tutor group compared to the SSR group (d = .35), as well as trends toward better improvement of fluency (d = .26) and overall reading comprehension (d = .28). The small size of the trend for higher fluency gains compared to the huge effect for word identification (d = 1.45) is puzzling, especially if it means that computer-guided practice in oral reading of connected text truly had much less impact on fluency than on word recognition. An alternative explanation is a difference in the measures used. For Word Identification, we used a nationally normed measure on an equal-interval W scale to combine data across grades 1-4. For fluency, we used an unnormed, curriculum-based measure of Words Correct Per Minute, which is known to increase sub-linearly with grade. Although sensitive to reading growth, this measure might conceivably under-estimate treatment effects when aggregating

fluency gains across grades, due to its non-equal-interval scale and to lower reliability than more carefully developed, published, normed measures.

Finally, although the effect sizes for spelling, fluency, and reading comprehension are considered "small" by Cohen's (1988) standards, a meta-analysis of effect sizes in the education literature found that the mean effect size for comparisons of educational interventions in elementary school students was d = .33, and the mean effect size in elementary schools when using standardized tests of narrow skills as the outcome, as we did, was even smaller (d = .23) (Hill et al., 2007). Therefore, the effect sizes we see are comparable to what can be expected when comparing educational interventions in elementary schools.

The pattern of effects shows that use of the Reading Tutor in comparison to sustained silent reading seems to have specifically helped with the identification of real words, perhaps by way of better-specified orthographic knowledge (evidenced by improvement in spelling) and the ability to piece together ordered component sounds of real words (evidenced by improvement on Blending Words). These skills are likely to have contributed to the trends we saw towards increases in both fluency and reading comprehension.

The skills that improved more for students using the computer reading tutor are the same skills specified by the National Reading Panel as improving from guided oral reading, namely "word recognition, fluency (speed and accuracy of oral reading), and comprehension" (NRP, 2000), and we additionally observed improvements in spelling. Thus, computerized guided oral reading based on empirical principles of good instruction can provide the benefits of guided oral reading while also providing the scalability of computer-based instruction. With oral reading, this benefit is especially notable, as allowing

students to individually read aloud for 20 minutes a day would require either more time or more instructors than are expected to be available in a typical classroom setting. The comparison to sustained silent reading shows that replacing guided oral reading with silent reading does not confer the same benefits.

Lastly, we mention the interaction between gender and treatment as a source of interest for future study. Although both males and females in the RT group performed better than SSR controls, there were trends towards significant interactions between treatment and gender on the Word Identification and Word Comprehension subtests of the WRMT, carrying over into the Basic Skills and Total Reading Composite scores. In these interactions, it appeared that males benefitted more from use of the Reading Tutor than the females. There are several caveats to interpreting this effect: the effect was small and only trending towards significance, and there is a known confound that males in the reading tutor group tended to be in higher grades than the females. In addition, the students in higher grades were selected to be among the poorer readers (in relatively good schools), whereas one of the first grade classes included all readers. Nevertheless, when considering the implementation of computerized instruction, it will be important to test and understand whether boys might benefit more than girls from technology-based education practices.

Why did the Reading Tutor improve reading more than Sustained Silent Reading? Candidate explanations explored in analyses of fine-grained data logged by the Reading Tutor (Mostow, 2004) include *motivation* to spend time productively (Beck, 2004, 2005, 2007; Mostow, Aist, et al., 2002; Mostow & Beck, 2003, 2007), *explicit instruction* such as the previews and reviews described in the Appendix, *scaffolding* on hard words and sentences (Beck et al., 2008; Heiner et al., 2004, 2005; Mostow, 2008; Mostow et al.,

2004), *transfer* to similar words (Leszczenski & Beck, 2007; Mostow, Beck, et al., 2008; Zhang et al., 2007), *harder text* than children choose, and *wide reading* of new text instead of rereading (Beck, 2006; Beck & Mostow, 2008).

Appendix

Types of Steps

Six types of steps comprised the activities and interventions in the 2000-2001 version of the Reading Tutor: Read, Listen, Spell, Pick, Edit, and Narrate.

Read steps consisted of assisted reading already described in the body of the paper.

In Listen steps, the Reading Tutor displayed text on the screen and read it aloud to the student. Such steps served to convey information or instructions too difficult for many students to read themselves. However, if they chose to read aloud, the Reading Tutor responded with the same assistance as in Read steps.

In Spell steps, the Reading Tutor prompted the student to spell a word aloud, turning the letters green after it heard them. This type of step was used in preview activities to practice new words.

In Pick steps, the Reading Tutor prompted the student to make a decision and displayed a large-font "talking menu" of alternatives to click on, highlighting and speaking each item in turn. For example, a Pick step from a story review activity let the student pick a "trouble word" to practice, from a menu of words that the Reading Tutor classified as misread during the story. The Reading Tutor read aloud the prompt at the top of the screen, and then each item on the list, highlighting it in yellow as it spoke. Besides implementing decision steps in activities, the Reading Tutor also used Pick steps to let students enroll themselves, log in, choose stories to read, and even let younger children create new stories by choosing a series

of words and phrases that the Reading Tutor then used to fill in blanks in a story template. Our analyses did not count this activity as writing.

In Edit steps, the Reading Tutor read a prompt aloud, and the student responded by typing in text using the keyboard. To reinforce spelling-to-sound mappings, the Reading Tutor sounded out each word as it was typed in, displayed a list of guesses for the intended word, and pronounced the word once completed. Writing activities at successively higher levels used Edit steps to input individual words, short answers, or entire stories.

In Narrate steps, the Reading Tutor prompted the student to read aloud a story s/he had written so as to add his or her voice to it. Narrate resembled Read but with fussier acceptance conditions, because its purpose was to capture an error-free narration of each sentence (Mostow & Aist, 1999a). Project LISTEN staff use Narrate to add their voices to stories (Mostow & Aist, 1999c).

Tutorial Interventions

To evaluate how well various tutorial interventions helped with different students, skills, and words, we embedded automated, within-subject, randomized-trial experiments in the Reading Tutor.

For example, we used the Word Identification Wrapper (described below) as an embedded experiment to evaluate alternative ways to preview new words before a story, based on 2,499 randomized trials administered and recorded by the 2000-2001 Reading Tutor (Mostow, 2008). The outcome variable was the student's ability to identify the new word correctly after the story, according to a human transcriber. A preview that required the student to spell the word aloud took longer but was significantly more effective than previews in which the Reading Tutor read the word aloud or also prompted

the student to echo the word. The oral spelling preview worked best – except for words over 9 letters long, for which it was not significantly better than having no preview at all.

A conventional between-subjects study tests the holistic effect of tutoring, as demonstrated by differences between treatment groups. In contrast, our embedded experiments test the individual effects of specific tutorial interventions. More broadly, they also show that tutoring affects student learning, by demonstrating the signature of tutorial decisions on subsequent student performance. However, they do not indicate how much those decisions affect overall reading growth. Future work needs to bridge the gap from such fine-grained evaluations to effects on gains in overall reading skills.

The 2000-2001 Reading Tutor selected among the following story previews, reviews, and "wrappers" (preview-review combinations) intended to exercise word identification, spelling, vocabulary, and comprehension, based on interventions that the National Reading Panel found effective (NRP, 2000):

- The New Word Simple Preview (all story levels) selected two new words from the story, displayed both words, and read them to the student.
- The New Word Type-in Preview (level B and above) additionally prompted the students to type in the two words.
- The Spelling Wrapper (level A and above) previewed only one of two new words, and prompted the student to spell it aloud. After the story, it posttested both words by prompting the student to type them in.
- The Clicked-Word Review (all levels) selected four words the student had clicked on in the story. To prompt the student to spell each word, it

displayed a sentence of the form "C A T spells CAT" as a Read step (unlike a Spell step, which spoke the word to spell but did not display it).

- The Trouble Word Review (all levels) selected four words the Reading Tutor had marked as misread in the story, and prompted the student to pick the hardest one. It then prompted the student to practice this word by typing it in.
- The Word Identification Wrapper (all levels) selected five new words from the story and introduced each one in a different way. After the story, it posttested each word by displaying it in isolation for the student to read aloud. Mostow (2008) reports the results of this experiment.
- The Word Comprehension Wrapper (level A and above) tested different ways to explain new vocabulary. It selected four new words for which it had story-specific definitions, and first asked whether the student knew what each word meant. To pretest a word, it presented the definition and asked the student to pick the matching word. It explained one word by giving its definition, explained another word by giving a synonym, pretested one word without explaining it, and left one word untested as a control. After the story, it posttested all four words using the same matching task as the pretest.
- The Poem Wrapper(all levels) asked the student to predict whether the poem would rhyme. After the poem, it asked if the prediction was correct, and if the student enjoyed the poem. The purpose of this and other prediction wrappers was to scaffold comprehension by activating relevant schemas before reading and stimulating reflection after reading. The National

Reading Panel (2000) found similar interventions by human teachers to be effective in developing comprehension. Tailoring such interventions to a specific text is easy for human teachers but hard to automate, so we designed our automated interventions to be generic and apply to a wide range of texts – for example, to any poem, as in the case of the Poem Wrapper.

- The Lower Level Fiction Wrapper applied to any narrative story at level B and below. It asked the student to think about the story's setting and lessons or ideas. After the story, it asked multiple-choice questions about who the story was about, whether the student would want to visit where it took place, and whether he or she learned anything new from the story.
- The Animal Story Review For Levels K, A, and B applied to animal stories. It prompted the student to type in the animal, and then asked multiple-choice questions about the animal's speed, height, whether it's a mammal, where it lives, and its desirability as a pet.
- The Animal Story Review For Levels C, D, and E prompted the student to type in both an animal from the story, and his or her favorite animal. Then it asked the student multiple choice questions about which animal is bigger, which animal is faster, the story animal's appearance and diet, and to type in where the story animal lives.
- The Main Character Wrapper applied to any fiction at level C or above. It asked the student to think about the main character's identity, problems, and solutions. After the story, it prompted the student to type in the character's

name and a problem. Then it asked whether the character overcame the problem, and whether the student would have done likewise.

- The Importance of a Title Wrapper (level C and above) prompted the student to type in what the story would be about. After the story, it read the prediction aloud and asked if the student had guessed right, and whether the title gave good clues about the story.
- The Story in Parts Wrapper (for sections of serialized fiction after the initial section) prompted the student to summarize the previous section aloud, then type in the name of the student's favorite character, a trait of that character, the setting of the story, and a prediction about what would happen next. After the story, it read the prediction aloud and asked if the student had been right.

Student Opinions

We also used the Reading Tutor to ask students' opinions. The Story Opinion Poll was one of the review activities the Reading Tutor chose among after the student finished reading any text. It asked the student to rate the interest, difficulty, and duration of the story just completed:

> What did you think of "the_story_title"? Click on your answer.
>
> It was fun. It was okay. It was boring.
>
> How hard was it? Click on your answer.
>
> It was too easy. It was okay. It was too hard.
>
> How long did it take? Click on your answer.
>
> It was too short. It was okay. It was too long.

The Reading Tutor administered its automated Story Opinion Poll after 4,972 completed stories, 2,471 chosen by the Reading Tutor and 2,501 by the student. The poll averaged 12 ± 9 seconds to administer. It showed that students preferred stories they picked, especially stories they reread. They rated as "fun" 29% of the stories the Reading Tutor picked, versus 36% of the new stories they picked themselves, and 41% of the stories they reread. Conversely, they rated as "boring" 36% of the stories the Reading Tutor picked, versus 31% of the new stories they picked themselves, and 29% of the stories they reread. Ratings of story length reflected similar preferences. Students rated as "too long" 37% of the Reading Tutor's choices, versus 28% of the new stories they picked, and 32% of the stories they reread. They rated as "too short" 31% of the Reading Tutor's choices, versus 38% of the new stories they picked, and 34% of the stories they reread. Interestingly, the percentages of stories rated as "too hard," "too easy," or "okay" in the Story Opinion Poll did not differ by more than 4% between stories chosen by the Reading Tutor, new stories chosen by students, and stories reread. Ratings of story enjoyment and length are a conservative indicator of preference in that they omit stories that students decided not to finish.

Besides polling students after each story they completed, we used the Reading Tutor near the end of the school year to administer an "end-of-year survey," implemented (like the introductory tutorials) as an activity chosen by the Reading Tutor for every student.

The survey included 18 multiple choice questions implemented as Pick steps, whose responses we tabulated automatically.

Eleven Pick steps asked about various genres, e.g.:

> How much do you like animal stories?

a lot a little bit not at all

Genres ranged in popularity (percentage of liked "a lot" responses) from mystery stories (67%), sports stories (59%), animal stories (57%), chapter stories (54%), science stories (43%), fairy tales (39%), and math stories (36%), down to history stories (30%), poems (28%), and letter and sound stories (20%).

Seven Pick steps surveyed students' attitudes toward reading and the Reading Tutor, e.g.:

How well does the Project LISTEN Reading Tutor help you read?

It does a GREAT job. It does an OKAY job. It does NOT help.

The percentage of students who responded "It does a GREAT job" decreased by grade, from 69% in grade 1 and 65% in grade 2 down to 33% in grade 3 and 27% in grade 4.

The survey also included 8 free-response follow-up questions, e.g.:

What do you think you've learned from using the Reading Tutor? Please answer OUT LOUD, then click Go.

We transcribed the recorded free-form spoken responses manually. Responses from children who rated the Reading Tutor as "great" in helping them read included:

Grade 1*: how to read; how to sound out; how to read read really hard words; I learned how to read really really good; I've learned to read better I've learned to write better I've learned to spell sentences better*

Grade 2: *you can always believe in yourself and sound out words and the the kyle the reading tutor will help you anytime you need*

Grade 3: *I learned bigger words; the reading tutor learned me how to read new words that I experienced*

Grade 4: *I learned a lot from the reading tutor because I found out new words that I haven't known*

Responses from children who rated the Reading Tutor OKAY included:

Grade 1: *that computers are the dumbest thing I've ever known in the whole universe that's what I learned; that the reading tutor I mean that reading on the computer can annoy you because it because it the computer always hates you*

Grade 3: *some science*

Grade 4: *how to say words if i don't know what they are*

Responses from children who said the Reading Tutor didn't help said they learned:

Grade 3: *some stuff about space and stuff; it did help me learn some stuff*

Grade 4: *couple words; nothing ... absolutely nothing*

Tables

Table 1: *Time allocation of all 88 students who used the Reading Tutor*

	Mean	SD	Min	Max
Total time	18:49:38	(3:38:57)	10:27:55	27:09:49
Reading	10:47:25	(2:46:10)	5:30:25	18:20:35
Previews and reviews	2:19:21	(0:48:44)	0:38:47	4:59:36
Writing	1:14:00	(0:59:03)	0:00:00	4:33:00
Percent time picking stories	5%	(2%)	1%	13%

Table 2: *Summary of Raw Gains*

	Sustained Silent Reading			Reading Tutor		
	Pre-test	Post-test	Gain	Pre-test	Post-test	Gain
CTOPP						
Blending Words	11.54 (3.46)	14.57 (3.43)	3.08 (3.86)	11.69 (3.48)	15.73 (2.91)	3.99 (3.15)
Elision	8.97 (4.96)	12.00 (4.97)	3.00 (4.79)	8.70 (4.68)	12.92 (4.31)	4.19 (4.36)
Rapid Letter	59.91 (23.92)	48.65 (11.83)	-11.66 (16.84)	59.15 (23.13)	46.81 (11.53)	-12.59 (15.88)
WRMT						
Word ID	427.84 (37.09)	439.18 (21.66)	11.33 (47.35)	429.50 (38.24)	468.89 (19.11)	39.39 (41.89)
Word Attack	470.06 (20.45)	487.32 (14.76)	17.27 (13.31)	470.32 (20.61)	486.72 (15.91)	16.40 (14.50)
Basic Skills	448.95 (28.06)	463.25 (12.21)	14.30 (27.54)	449.91 (28.49)	477.80 (12.82)	27.89 (24.55)
Word Comp	465.40 (22.61)	481.99 (16.00)	16.59 (10.90)	467.43 (24.05)	484.90 (15.64)	17.47 (12.32)
Passage Comp	460.49 (23.75)	477.46 (15.30)	16.96 (14.98)	461.08 (24.34)	479.53 (14.85)	18.45 (15.13)
Reading Comp	462.94 (22.54)	479.72 (15.19)	16.78 (11.41)	464.26 (23.56)	482.22 (14.86)	17.96 (12.26)
Total Reading	455.95 (24.96)	471.49 (11.47)	15.54 (18.13)	457.08 (25.68)	480.01 (11.91)	22.92 (17.30)
Oral Fluency	32.23 (28.98)	58.20 (32.18)	26.24 (16.53)	32.53 (30.57)	62.78 (34.20)	30.13 (19.31)
Spelling	5.96 (5.04)	11.01 (5.13)	5.24 (3.20)	5.73 (5.76)	11.89 (5.24)	6.50 (3.63)
ERAS	62.96 (10.63)	59.87 (13.10)	-2.74 (12.05)	60.26 (11.72)	57.70 (12.71)	-2.30 (12.18)

Table 3: *Summary of Treatment Effects*

	Adjusted SSR Gain	Adjusted RT Gain	Δ Adjusted Gain	Cohen's *d*	Partial η^2	*p*-value
CTOPP						
Blending Words	3.05 (.31)	4.04 (.31)	.99	.34	.02	.03
Elision	3.12 (.43)	4.07 (.43)	.95	.24	.02	.12
Rapid Letter*	-11.23 (.80)	-12.91 (.80)	-1.68	-.22	.01	.14
WRMT						
Word ID	39.53 (2.18)	10.29 (2.16)	29.24	1.45	.34	<.001
Word Attack	17.15 (1.12)	16.48 (1.13)	-.68	-.22	<.01	.671
Basic Skills	13.81 (1.23)	27.93 (1.24)	14.12	1.23	.27	<.001
Word Comp	16.02 (.76)	17.73 (.77)	1.70	.24	.01	.12
Passage Comp	16.64 (.97)	18.69 (.98)	2.05	.22	.01	.14
Reading Comp	16.31 (.73)	18.21 (.74)	1.90	0.28	.02	.07
Total Reading	15.04 (.77)	23.09 (.78)	8.06	1.11	.24	<.001
Oral Fluency	25.61 (1.93)	30.29 (1.94)	4.58	0.26	.02	.09
Spelling	5.24 (.36)	6.36 (.36)	1.13	0.35	.03	.03
ERAS	-2.86 (1.23)	-2.89 (1.27)	-.04	-.003	<.001	.98

*Note that for Rapid Letter Naming, a lower score is better.

References

Adams, M. J. (2006). The promise of automatic speech recognition for fostering literacy growth in children and adults. In M. McKenna, L. Labbo, R. Kieffer & D. Reinking (Eds.), *International Handbook of Literacy and Technology* (Vol. 2, pp. 109-128). Hillsdale, NJ: Lawrence Erlbaum Associates.

Adams, M. J. (2011). Technology for Developing Children's Language and Literacy: Bringing Speech Recognition to the Classroom. New York, NY: The Joan Ganz Cooney Center at Sesame Workshop.

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

Aist, G. S. & Mostow, J. (2008). Faster, better task choice in a reading tutor that listens. In V. M. Holland & F. P. Fisher (Eds.), *The Path of Speech Technologies in Computer Assisted Language Learning: From Research Toward Practice* (pp. 220-240). New York: Routledge.

Beck, J. E. (2004, August 31). *Using response times to model student disengagement.* In Proceedings of the ITS2004 Workshop on Social and Emotional Intelligence in Learning Environments, 13-20. Maceio, Brazil.

Beck, J. E. (2005, July 18-22). *Engagement tracing: using response times to model student disengagement.* In Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED 2005), 88-95. Amsterdam.

Beck, J. E. (2006, June 26). *Using learning decomposition to analyze student fluency development.* In ITS2006 Educational Data Mining Workshop, 21-28. Jhongli, Taiwan.

Beck, J. E. (2007, July 9-13). *Does learner control affect learning?* In Proceedings of the 13th International Conference on Artificial Intelligence in Education, 135-142. Los Angeles, CA.

Beck, J. E., Chang, K.-m., Mostow, J., & Corbett, A. (2008, June 23-27). *Does help help? Introducing the Bayesian Evaluation and Assessment methodology.* In 9th International Conference on Intelligent Tutoring Systems, 383-394. ITS2008 Best Paper Award. Montreal.

Beck, J. E., Jia, P., & Mostow, J. (2004). Automatically assessing oral reading fluency in a computer tutor that listens. *Technology, Instruction, Cognition and Learning, 2*(1-2), 61-81.

Beck, J. E., Jia, P., Sison, J., & Mostow, J. (2003, June 22-26). *Predicting student help-request behavior in an intelligent tutor for reading.* In Proceedings of the 9th International Conference on User Modeling, 303-312. Johnstown, PA.

Beck, J. E. & Mostow, J. (2008, June 23-27). *How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students [Best Paper Nominee]*. In 9th International Conference on Intelligent Tutoring Systems, 353-362. Montreal.

Campbell, R. (1988). *Hearing Children Read*. London and New York: Routledge.

Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). Effectiveness of Reading and Mathematics Software Products: Findings From Two Student Cohorts—

Executive Summary. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cunningham, T. (2006). *The Effect of Reading Remediation Software on the Language and Literacy Skill Development of ESL Students.* Master's thesis, University of Toronto, Toronto, Canada.

Hagen, A., Pellom, B., & Cole, R. (2007). Highly accurate children's speech recognition for interactive reading tutors using subword units. *Speech Communication, 49*(12), 861-873.

Heiner, C., Beck, J. E., & Mostow, J. (2004, June 17-19). *Improving the help selection policy in a Reading Tutor that listens.* In Proceedings of the InSTIL/ICALL Symposium on Natural Language Processing and Speech Technologies in Advanced Language Learning Systems, 195-198. Venice, Italy.

Heiner, C., Beck, J. E., & Mostow, J. (2005, July 18-22). *When do students interrupt help? Effects of time, help type, and individual differences.* In Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED 2005), 819-826. Amsterdam.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2007). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*, 172–177.

Jaffe, L. E. (2009). *Development, interpretation, and application of the W score and the relative proficiency index (Woodcock-Johnson III Assessment Service Bulletin No. 11).* Rolling Meadows, IL: Riverside Publishing.

Kantor, A., Cernak, M., Havelka, J., Huber, S., Kleindienst, J., & Gonzalez, D. B. (2012, September). *Reading Companion: The Technical and Social Design of an Automated Reading Tutor* In Workshop on Child, Computer and Interaction, Portland, Oregon.

Korsah, G. A., Mostow, J., Dias, M. B., Sweet, T. M., Belousov, S. M., Dias, M. F., & Gong, H. (2010). Improving Child Literacy in Africa: Experiments with an Automated Reading Tutor. *Information Technologies and International Development, 6*(2), 1-19.

Kuhn, M. R. & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology, 95*(1), 3–21.

Larsen, S. C., Hammill, D. D., & Moats, L. C. (1999). *Test of Written Spelling* (fourth ed.). Austin, Texas: Pro-Ed.

Lepper, M. R., Woolverton, M., Mumme, D. L., & Gurtner, J. (1993). Motivational Techniques of Expert Human Tutors: Lessons for the Design of Computer-Based Tutors. In S. P. Lajoie & S. J. Derry (Eds.), *Computers as Cognitive Tools* (pp. 75-105). Hillsdale, NJ: Erlbaum.

Leszczenski, J. M. & Beck, J. E. (2007, July 9). *What's in a word? Extending learning factors analysis to modeling reading transfer.* In Proceedings of the AIED2007 Workshop on Educational Data Mining, 31-39. Marina del Rey, CA.

McKenna, M. C., Kear, D. J., & Ellsworth, R. A. (1995). Children's attitudes toward reading: a national survey. *Reading Research Quarterly, 30*, 934-956.

Mostow, J. (2004, August 30). *Some useful design tactics for mining ITS data.* In Proceedings of the ITS2004 Workshop on Analyzing Student-Tutor Interaction Logs to Improve Educational Outcomes, 20-28. Maceió, Alagoas, Brazil.

Mostow, J. (2008). Experience from a Reading Tutor that listens: Evaluation purposes, excuses, and methods. In C. K. Kinzer & L. Verhoeven (Eds.), *Interactive literacy education: facilitating literacy environments through technology* (pp. 117-148). New York: Lawrence Erlbaum Associates, Taylor & Francis Group.

Mostow, J., Aist, G., Beck, J. E., 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?* In Proceedings of the Sixth International Conference on Intelligent Tutoring Systems (ITS'2002), 320-329. Biarritz, France.

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

Mostow, J. & Aist, G. S. (1997, July). *The sounds of silence: Towards automated evaluation of student learning in a Reading Tutor that listens.* In Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97), 355-361. Providence, RI.

Mostow, J. & Aist, G. S. (1999a, July). *Authoring new material in a reading tutor that listens.* In Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99), Intelligent Systems Demonstration track, 918-919. Orlando, FL.

Mostow, J. & Aist, G. S. (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.

Mostow, J. & Aist, G. S. (1999c). US Patent and Trademark Office.

Mostow, J. & Aist, G. S. (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. S., 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).* In Language Technologies 2001: The Second Meeting of the North American Chapter of the Association for Computational Linguistics., Pittsburgh, PA.

Mostow, J., Aist, G. S., Huang, C., Junker, B., Kennedy, R., Lan, H., Latimer, D., O'Connor, R., Tassone, R., Tobin, B., & Wierman, A. (2008). 4-Month evaluation of a learner-controlled Reading Tutor that listens. In V. M. Holland & F. P. Fisher (Eds.), *The Path of Speech Technologies in Computer Assisted Language Learning: From Research Toward Practice* (pp. 201-219). New York: Routledge.

Mostow, J. & Beck, J. E. (2003, November 3-4). *When the Rubber Meets the Road: Lessons from the In-School Adventures of an Automated Reading Tutor that Listens.* In Conceptualizing Scale-Up: Multidisciplinary Perspectives, Park Hyatt Hotel, Washington, D.C.

Mostow, J. & Beck, J. E. (2007). When the Rubber Meets the Road: Lessons from the In-School Adventures of an Automated Reading Tutor that Listens. In B. Schneider & S.-K. McDonald (Eds.), *Scale-Up in Education* (Vol. 2, pp. 183--200). Lanham, MD: Rowman & Littlefield Publishers.

Mostow, J., Beck, J. E., Chalasani, R., Cuneo, A., & Jia, P. (2002, October 14-16). *Viewing and analyzing multimodal human-computer tutorial dialogue: a database approach.* In Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces (ICMI 2002), 129-134. First presented June 124, 2002, at the ITS 2002 Workshop on Empirical Methods for Tutorial Dialogue Systems, San Sebastian, Spain. Pittsburgh, PA.

Mostow, J., Beck, J. E., & Heiner, C. (2004, June 27-30). *Which Help Helps? Effects of Various Types of Help on Word Learning in an Automated Reading Tutor that Listens.* In Eleventh Annual Meeting of the Society for the Scientific Study of Reading, Amsterdam, The Netherlands.

Mostow, J., Beck, J. E., Zhang, X., & Leszczenski, J. (2008, July 10-12). *Does fluency growth transfer among related words? Longitudinal evidence from Project LISTEN's Reading Tutor.* In Fifteenth Annual Meeting Society for the Scientific Study of Reading, Asheville, North Carolina.

Mostow, J., Roth, S. F., Hauptmann, A. G., & Kane, M. (1994, August). *A prototype reading coach that listens [AAAI-94 Outstanding Paper]*. In Proceedings of the Twelfth National Conference on Artificial Intelligence, 785-792. Seattle, WA.

NRP. (2000). Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Washington, DC: National Institute of Child Health & Human Development. At www.nichd.nih.gov/publications/nrppubskey.cfm.

Poulsen, R., Wiemer-Hastings, P., & Allbritton, D. (2007). Tutoring Bilingual Students with an Automated Reading Tutor That Listens. *Journal of Educational Computing Research, 36*(2), 191-221.

Reeder, K., Shapiro, J., & Wakefield, J. (2009, July 19-22). *A computer based reading tutor for young English language learners: recent research on proficiency gains and affective response.* In 16th European Conference on Reading and 1st Ibero-American Forum on Literacies, University of Minho, Campus de Gualtar, Braga, Portugal.

SAT-10. (2003). *Stanford Achievement Test, Tenth Edition (SAT-10), Intermediate 1: Directions for Administering (DFA)*. San Antonio, TX: Pearson Education, Inc.

SESAT. (1996). *Stanford Early School Achievement Test (SESAT), Level 2: Directions for Administering (DFA).* San Antonio, TX: Pearson Education, Inc.

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *The Comprehensive Test of Phonological Processing*. Austin, Texas: Pro-Ed.

Weber, F. & Bali, K. (2010, December 17-18). *Enhancing ESL Education in India with a Reading Tutor that Listens.* In Proceedings of the First ACM Symposium on Computing for Development 20:21-29. London, United Kingdom.

Williams, S. M., Fairweather, P. G., & Nix, D. (2008). Speech recognition to support early literacy. In C. K. Kinzer & L. Verhoeven (Eds.), *Interactive literacy*

education: facilitating literacy environments through technology (pp. 95-116). New York: Lawrence Erlbaum Associates, Taylor & Francis Group.

Wise, B., Cole, R., Vuuren, S. v., Schwartz, S., Snyder, L., Ngampatipatpong, N., Tuantranont, J., & Pellom, B. (2008). Learning to Read with a Virtual Tutor: Foundations to Literacy. In C. Kinzer & L. Verhoeven (Eds.), *Interactive Literacy Education: Facilitating Literacy Environments Through Technology*. New York: Lawrence Erlbaum Associates.

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

www.ed.gov. (2002, September 12, 2002). Blue Ribbons Schools—1982-2002 Retrieved December 4, 2002, from http://www.ed.gov/offices/OIIA/Recognition/nclb-brs/brs.html

Zhang, X., Mostow, J., & Beck, J. E. (2007, July 9). *All in the (word) family: Using learning decomposition to estimate transfer between skills in a Reading Tutor that listens.* In AIED2007 Educational Data Mining Workshop, Marina del Rey, CA.