A Controlled Evaluation of Computer- versus Human-assisted Oral Reading

Jack Mostow, Greg Aist, Paul Burkhead, Albert Corbett, Andrew Cuneo,
Susan Eitelman, Cathy Huang, Brian Junker, Cheryl Platz, Mary Beth Sklar, Brian Tobin
Project LISTEN, 4213 NSH, Carnegie Mellon University, Pittsburgh, PA 15213*
http://www.cs.cmu.edu/~listen

Abstract. A year-long study of 144 second and third graders compared outcomes (gains in test scores) and process variables (e.g. words read) for Project LISTEN's Reading Tutor, human tutors, and a classroom control. Human tutors beat the Reading Tutor only in word attack. Both beat the control in grade 3 word comprehension.

1. Experimental Design

Project LISTEN's Reading Tutor listens to children read aloud and helps them [1], and also lets them write and narrate stories [2]. In 1999-2000, to "prove and improve" the Reading Tutor – to evaluate against conventional instruction, and identify areas for improvement – we compared it to one-on-one human tutoring, and to spending the same time in class.

Students. Students were 144 second and third graders (ages 7-10) at an urban elementary school near Pittsburgh, Pennsylvania, that had not previously used the Reading Tutor. Teachers in 12 classrooms each chose their 12 poorest readers, based on a finding that the 1998 Reading Tutor seemed to make a bigger difference for students in the bottom half [3].

Treatments. We assigned each student to the same treatment 20 minutes daily for the year. Each day included 60-70 minutes of reading instruction plus varying time on related activities. Thus all students in a given classroom received mostly the same instruction, with tutored students averaging 0 to 15 minutes more time per day on reading and writing. Teachers rotated scheduling to vary which subjects students missed while being tutored.

Control. Regular instruction used a basal reading curriculum, with class size about 24.

Reading Tutor. Children took turns throughout the school day using one Reading Tutor computer in their classroom. Teacher cooperation was essential to this arrangement, so the principal chose six classrooms to get Reading Tutors based on his estimate of teachers' willingness to cooperate – possibly a confound, but necessary. Moreover, according to the principal, all classroom teachers in the study were comparably experienced veteran teachers.

Human tutoring. Variables in tutoring include *personnel*, *setting*, *activities*, and *materials*.

Personnel: 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 [4] to paraprofessional teachers' aides to certified teachers to certified teachers with specialized training in a particular reading program. Using certified teachers rather than paraprofessionals has been associated with positive results for one-on-one reading tutoring [5]. 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

* This work was supported in part by the National Science Foundation under Grant No. REC-9979894, and by Greg Aist's NSF Graduate Fellowship and Harvey Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this publication are the authors' and do not necessarily reflect the views of the National Science Foundation or the official policies, either expressed or implied, of the sponsors or of the United States Government. We also thank Centennial Elementary students and educators, and the reviewers.

to do better than classroom instruction, but not as well as the world's best tutor – an unrealistic comparison even for a research study, let alone for large-scale implementation.

Setting: Each tutor tutored 6 students from one class, one at a time at a desk in the hall.

Activities: Tutors helped students read and write, and logged each session's activities.

Materials: To control for materials, we asked human tutors to use bound copies of the same stories used in 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.

Assignment of students to treatments. We initially assigned 60 students to use the Reading Tutor, 36 students to human tutors, and 48 students to the control condition. To keep either type of tutoring from influencing the other, each classroom had only one type.

Ten students was the maximum we thought could share one Reading Tutor. Two teachers tried to add 1-2 more students, but could not always get them on. We excluded these 3 "part-timers" from analysis. Other students in each Reading Tutor room were in-room controls.

Similarly, 6 students was the most each human tutor could cover, given her other duties. The other 6 students in the same room served as in-room controls, likewise chosen so as to make treatment groups statistically well-matched. 131 of 144 students completed the study.

Outcome measures. The Woodcock Reading Mastery Test (WRMT) [6] is an individually administered test normed by month within grade, so that a pre- to post-test gain of 0 means the student stayed at the same percentile relative to peers. WRMT subtests measure different reading skills with mean 100 and standard deviation 15. Trained testers pre-tested students in September 1999 and post-tested them in May 2000, using four WRMT subtests: Word Attack (WA) for decoding skills, Word Identification (WI) for reading single words, Word Comprehension (WC) for word meaning, and Passage Comprehension (PC) for understanding text. The testers also measured students' independent oral reading fluency as the median number of words read correctly in one minute for three grade-level passages.

2. Results

We found surprisingly few differences among treatments. We expected the human tutors to lead across the board. Instead, as Table 1 shows, human tutoring significantly outgained the Reading Tutor only in Word Attack. Human and computer tutoring both surpassed the control in grade 3 Word Comprehension gains. No other differences were significant. In grade 3 Passage Comprehension, a trend favored the Reading Tutor over the control. The absence of significant differences in fluency gains is especially surprising, because fluency is such a sensitive measure of growth [7]. A few differences among tutors were significant [8].

Table 2 compares Reading Tutor and human tutor process variables, using comprehensive records at multiple levels of detail, plus hand-coded videotapes of 49 sample sessions.

Table 1: Comparison of treatment groups' pretest scores and gains on each test, by grade

Subtest	**Pretest Score**			**Gains**			**ANOVA**	**Main**	**Contrast**
	CTRL	**RT**	**HT**	**CTRL**	**RT**	**HT**	**covariates**	**Effects**	**Effect**
Grade 2:	**n=19**	**n=29**	**n=17**				**(pretests)**	**p**	**Size**
Word Attack (normed)	86.6	84.6	83.3	8.2	**3.1** <?	**11.0**	WI, WA	0.07	**0.58**
Word ID (normed)	90.2	90.3	89.3	1.6	-0.7	1.0	WI, WC	0.16	
Word Comp (normed)	89.7	88.4	90.4	5.6	4.4	4.4	WI, WC	0.73	
Passage Comp (normed)	90.8	88.9	89.5	1.9	2.3	2.0	WC, PC	0.90	
Fluency (WPM)	12.9	12.9	15.4	39.1	34.8	40.7	FLU	0.67	
Grade 3:	**n=20**	**n=29**	**n=17**						
Word Attack (normed)	93.4	95.3	93.7	-1.1	**-2.8** <?	**3.6**	WI, WA	0.10	**0.65**
Word ID (normed)	90.1	90.3	91.3	0.3	1.8	1.8	WI, WC	0.53	
Word Comp (normed)	89.7	90.4	94.3	**0.7** <<	**4.3**	3.4	WI, WC	0.02	**0.56**
Passage Comp (normed)	89.2	89.5	90.9	**1.0** <?	**4.8**	4.1	WC, PC	0.14	**0.48**
Fluency (WPM)	42.0	37.7	41.8	19.9	20.5	28.1	FLU	0.18	

Table 2: Comparison of process variables for Reading Tutor (RT) and human tutoring (HT), by grade

Process variable, data source (and how derived)	**Grade 2**		**Grade 3**	
[averaged by student or per videotaped session;	**Reading**	**Human**	**Reading**	**Human**
shown by grade and by RT room or HT initials]	**Tutor** n=29	**tutor** n=17	**Tutor** n=29	**tutor** n=17
Total number of sessions	**67 days**	**73 days**	**71 days**	**>>61 days**
RT event database (days with any events)	90 RT201	67 AC	70 RT301	61 LN
HT log (days with any logged activity)	54 RT211	77 MB	57 RT303	62 MM
	56 RT212	77 ME	86 RT304	58 NJ
Reading/total time in videotapes (hand-coded; other time included writing, waiting for RT, etc.)	**11/20 min.**	**11/18 min.**	**9/20 min.**	**7/15 min.**
Story words seen per session	**122 words**	**<? 154 words**	**143 words**	**<< 262 words**
RT portfolio (#words of finished stories only!)	120 RT201	112 AC	122 RT301	258 LN
HT log (#words in logged stories; prorated for	108 RT211	224 MB	143 RT303	313 MM
never-finished stories based on # pages read)	135 RT212	120 ME	162 RT304	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;	1.1 RT201	1.4 AC	1.4 RT301	2.3 LN
finished stories averaged a half level lower.)	0.8 RT211	2.8 MB	2.0 RT303	2.2 MM
HT log (shows level, pages read, not who chose)	1.2 RT212	1.2 ME	1.7 RT304	2.2 NJ
Percentage of rereading	**30%**	**>> 19%**	**24%**	**>> 13%**
RT portfolio (% of finished stories read before)	34% RT201	24% AC	25% RT301	13% LN
HT log (% of finished stories read before)	28% RT211	11% MB	18% RT303	18% MM
	29% RT212	21% ME	30% RT304	6% NJ
Percent of sessions with any writing activity	**38%**	**<< 64%**	**28%**	**<< 58%**
RT event logs (% of days with edit events)	46% RT201	85% AC	36% RT301	67% LN
HT log (listed writing activities)	37% RT211	37% MB	25% RT303	60% MM
	32% RT212	70% ME	22% RT304	44% NJ

Control. A questionnaire asked each teacher how much time her class spent on reading.

Reading Tutor. The Reading Tutor recorded student utterances, speech recognizer output, a detailed event log, a more selective event database for runtime reference, student portfolios listing stories read and new words seen, and counts of distinct words and stories read.

Human tutoring. We analyzed tutors' session logs, and archived students' writing journals.

This diverse data may help us explain outcome differences and improve the Reading Tutor.

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

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

[2] J. Mostow, G. Aist, C. Huang, B. Junker, R. Kennedy, H. Lan, D. Latimer, R. O'Connor, R. Tassone, B. Tobin, and A. Wierman. 4-month evaluation of a learner-controlled reading tutor that listens. In Philippe DeCloque and Melissa Holland (Editors), *Speech Technology for Language Learning*. The Netherlands: Swets & Zeitlinger Publishers. In press, 2001.

[3] J. Mostow and G. Aist. Authoring new material in a reading tutor that listens. *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, Orlando, FL, July 1999, 918-919.

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

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

[6] American Guidance Service. Bibliography for Woodcock Reading Mastery Tests – Revised (WRMT-R). http://www.agsnet.com/Bibliography/WRMTRbio.html

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

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