7

When the Rubber Meets the Road: Lessons from the In-School Adventures of an Automated Reading Tutor That Listens[1]

Jack Mostow and Joseph Beck

In principle, technology has the potential to scale up educational interventions by automating them. Automated intervention ensures treatment fidelity over time and across multiple schools and settings in a way that human instruction cannot. Automated tutoring is consistent whether studied in small- or large-scale implementations—a key property for scalability. That is, contextual factors may affect the amount of tutoring (*implementation intensity*), but not the intervention itself. This point is a bit more subtle than it may seem, because automated interventions can and do respond differentially to student behavior and teacher input. However, the intervention *design* remains intact.

In contrast, human interventions can be transformed when implemented (Coburn 2001; Elmore 1996; Hoffman, McCarthey, and Elliot et al. 1998; Spillane and Jennings 1997; Spillane and Zeuli 1999; Stein, Grover, and Henningsen 1996), with teachers altering interventions in ways that can negate the very components responsible for strong effects in experimental trials. Lasting educational change is difficult to achieve because it must change not only educational materials but teacher practice and beliefs (Fullan 2001). Other projects funded by IERI (IERI 2002) and the U.S. Department of Education focus on trying to fundamentally change the ways teachers teach reading. However, the more a new practice differs from an existing practice, the harder it is to implement (Cohen and Ball 2001). The main changes required to implement an automated intervention are to schedule and value its usage. These requirements are challenging, but straightforward compared to deeper transformations in how teachers teach. Ideally, the technology is installed, students use it, and they learn.

In practice, the scalability of a technology-enabled intervention depends on how it is implemented in schools (Cuban 2001; Schofield 1995; Schofield and Davidson 2002). Project staff, requirements, teachers, technology, and student factors all affect learning (Steuck, Meyer, and Kretschmer 2001). For example, in reality the technology is installed, (some) students use it, and it breaks. What happens next is one of many crucial implementation factors.

The theme of this chapter is the use of technology not only to automate an intervention but to help analyze its actual implementation. We describe what we are learning about scalability in the context of a particular technology-enabled intervention—Project LISTEN's automated Reading Tutor, which we now describe.

PROJECT LISTEN'S READING TUTOR

Project LISTEN's Reading Tutor displays text on the computer screen and *listens* to a child read it aloud (Mostow and Aist 1999, 2001) using an inexpensive noise-canceling headset microphone and a Windows™ personal computer. The Reading Tutor adapts the Sphinx-II speech recognizer (Ravishankar 1996) to analyze children's oral reading (Aist 1999; Aist, Chan, and Huang et al. 1998; Aist and Mostow 1997a, 1997c; Fogarty et al. 2001; Hauptmann, Chase, and Mostow 1993; Mostow and Aist 1997; Mostow, Beck, and Winter et al. 2002; Mostow, Hauptmann, and Chase et al. 1993; Mostow, Roth, and Hauptmann et al. 1994). The Reading Tutor gives spoken and graphical help, shown at www.cs.cmu.edu/~listen/mm.html in a three-minute PBS video clip (Rubin 2002), and is based on effective human interventions (NRP 2000; Snow, Burns, and Griffin 1998).

Student Interaction with the Reading Tutor

A Reading Tutor session starts when a student clicks the *Hello* icon and logs in by selecting his or her name from a talking menu. The session ends when the student clicks *Goodbye* or the Reading Tutor times out after prolonged inactivity. To help teachers manage usage, the Reading Tutor displays a roster between sessions to show who has read that day, and for how long.

During a session, the student and Reading Tutor take turns picking which activity to do next (Aist 2000b; Aist and Mostow 2000, forthcoming). To help teachers monitor use, a status window at the top of the screen shows which student is logged in, how long the student has been on, the student's level, the title and level of the activity, and how many times the student has completed that activity before, as figure 7.1 illustrates.

An activity consists of one or more steps of a few types: assisted reading; listening to the Reading Tutor read; writing, including typed spelling; and

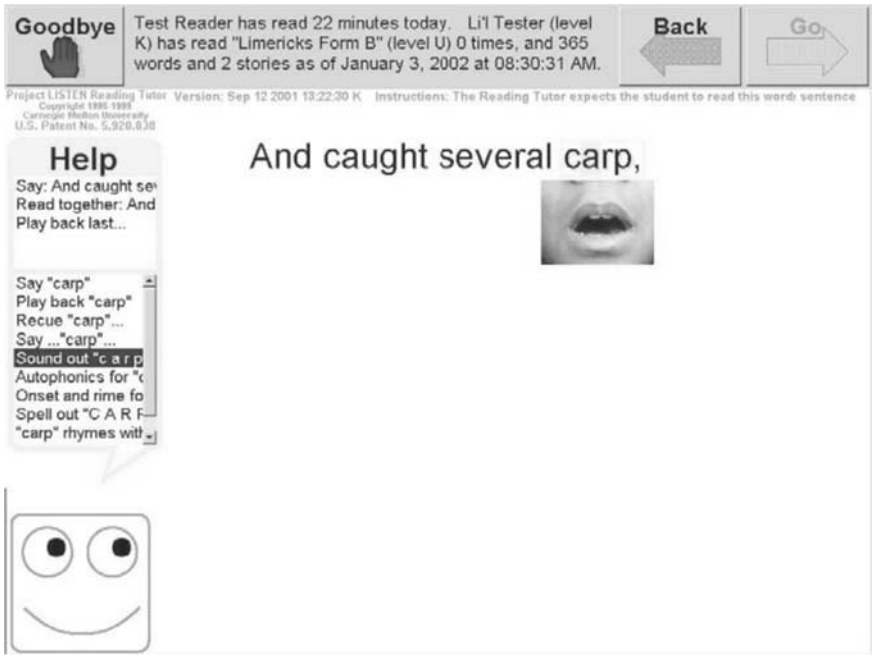

Figure 7.1. Screen Shot from Project LISTEN's Reading Tutor
Aist et al. 2002.
The Reading Tutor is in the middle of sounding out *carp*.

making a choice by selecting from multiple items in a talking menu. In assisted reading, the Reading Tutor displays text incrementally, adding a sentence or phrase at a time. It listens to the student read aloud, shadowing the next word and making its displayed persona gaze at it. The Reading Tutor responds verbally and/or graphically when it notices the student get stuck, skip a word, click for help, or finish the sentence. The Reading Tutor may intervene before, during, and/or after an activity, for example, to explain a new word, to give a hint on a hard word, to read a difficult sentence aloud, or to insert a comprehension question.

Teacher Interaction with the Reading Tutor

Other than the information displays mentioned above and the reports described below, direct teacher interaction with the Reading Tutor is limited by design, because most teachers have little time or inclination to use it themselves outside of initial training. For example, we designed the enrollment process to let students enroll themselves with little or no teacher help. However, the Reading Tutor does have a few password-protected teacher-only operations. To keep students from enrolling multiple times

under different names (as some of them liked to do), we password protected the first enrollment of the day, leaving enrollment open the rest of the day. This solution let students enroll themselves at the start of the year with minimal burden on the teacher, but prevented subsequent reenrollment of the recreational kind.

Exiting the Reading Tutor is password protected, except on school-owned computers where students must exit after every session. A password-protected *Fix* menu lets teachers adjust the level of or hide any student or story—hide rather than delete, so as to limit the damage in case a mischievous student should learn the password.

Automated Assessment

We use data captured by the Reading Tutor to generate automated continuous assessments of students' reading skills, in contrast to time-consuming individual tests that teachers can administer only occasionally and that interrupt instruction—much like closing down a store to take inventory of its stock (Pellegrino, Chudowsky, and Glaser 2001, 284). Data used for assessment include the latencies preceding each word the student reads aloud (Beck, Jia, and Mostow 2003, 2004; Jia, Beck, and Mostow 2002; Mostow and Aist 1997), students' help requests (Beck, Jia, and Sison et al. 2003), and multiple-choice fill-in-the-blank comprehension questions with automatic generation, scoring, and instant feedback (Mostow, Tobin, and Cuneo 2002).

There are multiple audiences for these assessments. The Reading Tutor uses its continuous assessment to adjust the level of stories chosen and help given. Reports on class and individual student usage and progress are generated on demand from the database to provide information that teachers want (Alpern et al. 2001). A password-protected website lets teachers obtain up-to-date reports anytime from school or home. The database also records which report(s) each teacher chooses, so that we as researchers can analyze which information teachers actually use, and how it relates to student usage.

SCALING UP THE READING TUTOR, 1996–2003

Deployment of the Reading Tutor has scaled up along several dimensions from a 1996–1997 pilot study (Aist and Mostow 1997b; Mostow 1997) to field tests in classrooms (Mostow 1998) to daily use in 2003 on nearly 200 computers at nine schools by 600 K–4 students spanning a wide range of reading abilities. Data logged by 176 of the Reading Tutor computers during the 2002–2003 school year showed that they provided 595 readers with

26,362 tutoring sessions totaling 3,839 hours, 38 minutes, and 59 seconds. The Reading Tutors listened to students read millions of words, and answered hundreds of thousands of requests for help on difficult words and sentences.

Implementation

Population

The number of sites has increased from a single school to nine diverse schools in six districts in two states, including urban and suburban locations, low-income and affluent communities, and African American and white students. The number of students in our studies rose from eight to several hundred. Their grade levels expanded from grade three down to kindergarten and as high as grade nine, though most were in grades one through four. The number of teachers has increased approximately tenfold from the four teachers of the eight students in the 1996 pilot.

Technical Issues

We made the transition—crucial for scalability—from running only on computers we owned and controlled (of which we purchased over 100) to running on school-owned Windows™ (2000 or XP) computers, which now account for over half the computers running the Reading Tutor. The installation process still has a few cumbersome manual steps, but has largely been automated using InstallShield™ from a set of compact disks, and/or Ghost Installer™ to "clone" an already configured hard disk. Harvesting research data from schools has progressed from manually "milking the machines" to automatically sending data via Internet to our lab (except for comprehensive recordings of oral reading, which we still harvest by hand due to Internet bandwidth restrictions). The configuration of the Reading Tutor has advanced from a stand-alone version that required students to use the same computer they originally enrolled on to a client-server configuration that lets students use any Reading Tutor in their school, and provides teachers with password-protected web-based reports they can access from any browser, including at home.

Supervision

The settings have scaled up from individually supervised pullout to independent use in school labs, classrooms, specialists' offices, and resource rooms. Training students to operate the Reading Tutor—a process originally performed by field support staff—is now handled by automated interactive

tutorials automatically presented to new users. Manual assessment and leveling have given way to automated versions.

Design Iteration

Scaling up to robust use in real schools and new settings poses design challenges that require considerable iteration to identify and address, as the following anecdotes illustrate.

When we first scaled the Reading Tutor from individually supervised use to a summer reading lab with eight computers in one room, we experienced a wave of mystery "crashes" that turned out to consist of students exiting the Reading Tutor by clicking on the X in the upper right corner of the window when the lab monitor was not looking. We redesigned the interface to eliminate this method for telling Windows™ to close the window.

The Reading Tutor originally let students freely click *Go* to advance to the next sentence. When students used the Reading Tutor side by side, some of them raced through stories by clicking *Go* without reading. We modified the Reading Tutor to enable *Go* only after it heard the student read at least half the words in the sentence.

The Reading Tutor had displayed separate lists of old and new stories to read. The students interpreted the number of old stories as a score, which is what motivated the racing behavior. We made sure that if the Reading Tutor displayed anything interpretable as a score, it would encourage educationally productive behavior. For example, we displayed the number of distinct words seen, rather than the total number of words seen, so as to encourage students to pick new, challenging stories, rather than reread the same easy stories to rack up easy "points."

Some Reading Tutor stories caused unforeseen problems in classrooms. A story adapted from *Weekly Reader* about Mighty Morphin' Power Rangers caused rambunctious behavior in at least one classroom, leading the teacher to hide the story using the password-protected feature described above. A similar problem involved one of the kindergarten-level stories we wrote about different letters of the alphabet and words that start with them. For example, "The Letter A" started out "APPLE starts with A," with a picture of an apple. We had to hide "The Letter H" because its picture of a farm implement failed to prevent an unanticipated interpretation of the example word HOE.

Relation among Context, Usage, and Impact

Figure 7.2 summarizes our model of the relationship among the context in which the Reading Tutor is used, how much it is used, and its impact on student achievement. We seek to evaluate and improve both usage and efficacy.

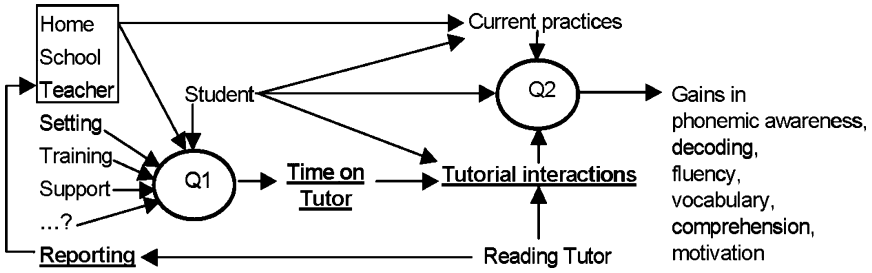

Figure 7.2. Reading Tutor in Context
Underlined items are instrumented.

Usage—how much a given student uses the Reading Tutor—depends on student, home, school, and teacher variables, such as teacher perceptions of the Reading Tutor's utility. Factors specific to the Reading Tutor include setting (number and location of computers), professional development and technical support, and reports to teachers, all of which may affect teacher perceptions.

The Reading Tutor's efficacy is defined by its impact on achievement gains for a given amount of usage. This impact depends not only on how much the Reading Tutor is used, but also on the current practices that it replaces or augments. Current practices include the reading program embodied in the curriculum, the text materials, the teaching methods employed, and the schedule of instruction, such as the amount of time devoted to language arts. Thus its impact should not be measured simply by students' gains in reading, but by how much *greater* their gains are than they would be otherwise, based on similar students who do not use it, but who start with similar skills and receive similar classroom instruction (Mostow, Aist, and Burkhead et al. 2003; Mostow and Beck, forthcoming).

Analyzing Usage

What might influence usage, whether directly or indirectly? Students' attitudes and attendance may affect how much they use the Reading Tutor. Teachers' attitudes, beliefs, scheduling practices, and organizational skills may affect their students' usage. Parental expectations, district policies, school schedules, and principals' attitudes may all affect usage.

The Reading Tutor's setting in classrooms, labs, or specialists' offices affects costs, convenience, and usage through its effects on ease of scheduling and degree of supervision. A cluster of two to four computers in a classroom can be easier for a teacher to schedule than a single computer that only one student can use at a time. A lab that students leave their classroom to use can reduce the burden on the classroom teacher to monitor Reading Tutor

use—at the costs of finding space (a scarce resource in most schools) for the lab, staffing the lab, coordinating its schedule with other teachers, and moving Reading Tutor use to where classroom teachers may not see it. A lab large enough for the whole class frees the teacher to monitor Reading Tutor use without trying to teach a lesson at the same time.

The technical environment may influence usage. The reliability of the Reading Tutor software and hardware influences the frequency of technical glitches. Software usability, teacher training, and technical support affect how long it takes to recover from such glitches—minutes if someone in the room knows how (for example by simply relaunching the program), much longer otherwise (for example to diagnose a broken microphone if the interface does not make it obvious).

Information from and about the Reading Tutor may influence usage by affecting teacher perceptions. If the Reading Tutor's assessments of student usage and progress are useful to teachers, teachers may (we hope) make sure their students use it. Conversely, if they see Reading Tutor behavior whose purpose they do not understand, they may be reluctant to put students on it.

Instrumenting Usage

How can we measure these various influences on usage? To be ecologically valid, measurement of student and teacher behavior must not only occur in normal settings, it must be unobtrusive (Webb et al. 1973). We use overt methods such as interviews and live observation to provide invaluable qualitative information, but they can distort classroom behavior. For example, teachers who normally have low usage will typically put students on the Reading Tutor when we visit.

We address this problem by using exquisitely detailed instrumentation to study scalability issues. Starting with the 2002–2003 version, the Reading Tutor logs every interaction with every student directly into a database, and sends these data overnight via Internet to our lab. Previous versions logged detailed information to text files that were laborious to collect and unwieldy to parse. Now our database server at each site simply sends back all its transactions. The resulting aggregated database (Mostow, Beck, and Chalasani et al. 2002) makes it possible to measure Reading Tutor usage precisely and comprehensively, analyze how it varies with context, and relate it to student progress. Compared to conventional methods for estimating implementation fidelity and intensity, this approach scales up data collection by orders of magnitude in multiple dimensions—to *all* students and teachers who use the Reading Tutor (not just a few); continuous, longitudinal measurement (not just occasional visits); and copious levels of detail (not just samples and summaries).

To take a simple but real example, Reading Tutor reliability can be quantified by the percentage of sessions that end in crashes. A quick query to our aggregated database generated a table showing how this percentage varied day by day as we deployed and debugged the fall 2002 version of the Reading Tutor. We could similarly quantify how long it takes to recover when a crash occurs—which may influence not just the Reading Tutor's availability, but also teacher acceptance. Dependability is crucial to the use of technological innovations like computers in schools (Davidson, Schofield, and Stockes 2001).

Quantifying Usage

A basic question pertaining to sustainability and scalability is how usage is affected by the various contextual variables shown in figure 7.2. Is usage greater or more consistent in some settings than others? How much does usage vary by student, teacher, grade, setting, and school? How much variance do each of these variables contribute?

We can measure student usage and quantify how it varies from room to room. For example, analysis of usage data from Reading Tutors in six classrooms at one elementary school in 2001–2002 showed that total usage for the school year varied considerably, ranging from a per-student average of 13.02 ± 8.02 hours in one room to 26.14 ± 7.09 hours in another. Differences in average usage between rooms at the same school reflect effects of teacher and grade. Usage varied within rooms as well, with standard deviations ranging from 2.27 to 8.02 hours, or 12 percent to 62 percent of the class mean; yet classroom accounted for almost half the total variance in usage (adjusted $R^2 = 0.426$, $p < 0.001$). Differences in average usage between students in the same classroom reflect student effects, such as attendance and motivation. Such differences may also reflect teacher effects, such as policy for which students to put on the Reading Tutor. The within-room standard deviation quantifies the amount of variation in usage by students within a classroom. This quantity reflects the heterogeneity of the students in the classroom, but also the teacher's policy and classroom management style. A small standard deviation indicates that the teacher made sure that all students spent approximately the same time on the Reading Tutor. A large standard deviation might indicate that the teacher deliberately assigned more time to some students than others. For example, a teacher might give weaker readers more time on the Reading Tutor, or might use extra Reading Tutor time to reward good behavior. Alternatively, a large standard deviation might indicate that the students in that classroom determined how much time they spent on the Reading Tutor. That is, the standard deviation might quantify how much slack the teacher gave students to determine their own time allocation.

Session *frequency* and session *duration* are different phenomena. Based on previous anecdotal evidence, we hypothesized that teachers would determine session frequency by deciding when to allow Reading Tutor usage, but students would control session duration by deciding how long to stay on. Class averages for session frequency and duration ranged from 1.53 ± 0.65 to 3.21 ± 0.37 days/week (compared to a nominal target of 4 days/week) and from 17.91 ± 1.40 to 21.73 ± 4.93 minutes/day (with 20 minutes/day as the nominal target). Classroom accounted for *most* of the variance in frequency (adjusted $R^2 = 0.678$, $p < 0.001$), but for *none* of the variance in duration (adjusted $R^2 = -0.004$, $p = 0.5$), strongly supporting our hypothesis, at least for this particular school and year. That is, teachers decided when to put students on, and students decided when to get off.

Quantifying Influences on Usage

The database introduced in the fall 2002 version of the Reading Tutor made possible a more refined analysis of usage. In the 2002–2003 school year, we had four types of settings:

- **Classroom**: Students took turns using one or more Reading Tutors in their classroom while other students received regular instruction.
- **Whole-class lab**: Students used the Reading Tutor outside their regular classroom in a lab equipped with some number of computers. A "whole-class lab" has enough computers that the teacher can take the entire class to go use the Reading Tutor at the same time. In previous years some schools had supervised "pullout labs" where students went to use the Reading Tutor while the teacher continued to teach the rest of the class. For the 2002–2003 school year, all the labs were whole-class labs.
- **Specialist**: Some students used the Reading Tutor in a reading or learning specialist's office equipped with one to three computers. Some of these students also used the Reading Tutor in a lab or classroom.
- **Resource**: One school had five Reading Tutors in the library and a sixth in a lab where the rest of the computers did not run the Reading Tutor. All six of these Reading Tutors were used to accommodate occasional overflow from classroom Reading Tutors, so we analyzed their usage as part of classroom usage, even though they were shared among classes.

Table 7.1 shows the distribution across grades and settings of the 396 students who used the Reading Tutor for a total of at least one hour during the 2002–2003 school year. Six students used the Reading Tutor in more than one setting and are counted separately for each setting. We focus on the 350 students who were in grades one through three.

Table 7.1. 2002–2003 Usage Data by Setting and Grade

Setting:	*K*	*1*	*2*	*3*	*4*	*5*	*6*
Class	15	**52**	**75**	**40**	20		
Lab		**92**	**43**	**43**			
Specialist		**1**	**4**		2	2	7

Note: Cell values are numbers of students. Most were in grades 1–3.

To analyze usage, we defined two outcome variables. *Session frequency* describes how often a given student used the Reading Tutor. To quantify session frequency meaningfully, we must specify over what time interval, and relative to what target. We deployed the Reading Tutor on different dates in different schools, and some students joined or left a class in the middle of the year. Therefore, we define the time interval for a given student as starting on the first calendar day when that student used the Reading Tutor, and ending on the last such calendar day. To exclude weekends, holidays, snow days, and so forth, we express session frequency as a percentage of the number of days the student could possibly have used the Reading Tutor. To exclude bogus dates caused by students' resetting the date on some computers, we define a possible day of usage as one when at least five students used the Reading Tutor anywhere—not necessarily the same site. The resulting session frequencies are therefore diluted by student absenteeism, and by assemblies that precluded usage at one school but not another. We average session duration for a given student over the days when that student used the Reading Tutor.

We performed some statistical analyses to quantify the magnitude of various contextual influences on usage:

- **Setting**: lab, classroom, specialist's room, or resource room
- **School**: identity of the particular school of the eight schools
- **Teacher**: individual identity among twenty-three teachers in eight schools (twelve teachers who used the Reading Tutor in their classrooms, nine teachers who used labs, and two specialists)
- **Grade**: one, two, or three in our subsample of 350 students
- **Ability**: We subtracted students' grade from the average of their pre- and posttest grade-equivalent Total Reading Composite scores (Woodcock 1998) to estimate how far above or below grade level they read.
- **Computer-student ratio**: number of Reading Tutors in a room, divided by the number of students using the Reading Tutor in that room

Many of these influences are conflated. For example, each teacher taught at only one school and used only one setting. Therefore we had to analyze some influences separately.

Table 7.2. Mean Session Frequency and Duration, by Setting

	Lab	*Class*	*Specialist*
Frequency	40.9%*	31.4%*	16.0%
Duration	18.9*	15.1	13.4

Note: * indicates statistically significant difference ($p < .001$ for lab versus class)
Means are adjusted to control for differences in grade and ability.

As table 7.2 shows, both the number and duration of sessions, averaged per student, turned out to be significantly higher for labs than for classroom settings. These quantities were lower for specialist settings. To level the playing field, this comparison adjusts usage for differences in student grade and ability. The adjusted estimates are for grade = 1.81 and ability = 0.85 above grade. Grade is an integer, but grade level increases from, say, 1.1 to 1.9 as the school year progresses, so 0.85 actually means only about a third of a year higher than where the student should be on average.

Figure 7.3 compares lab and classroom usage by room. Here usage is expressed as the per-student average number of minutes per day, *not* per session, so as to combine session frequency and duration into a single composite measure of average individual usage. Figure 7.3 illustrates some important points. First, usage averaged far below our nominal target of 20 minutes per day. Second, usage averaged 7.96 ± 2.96 minutes per day for

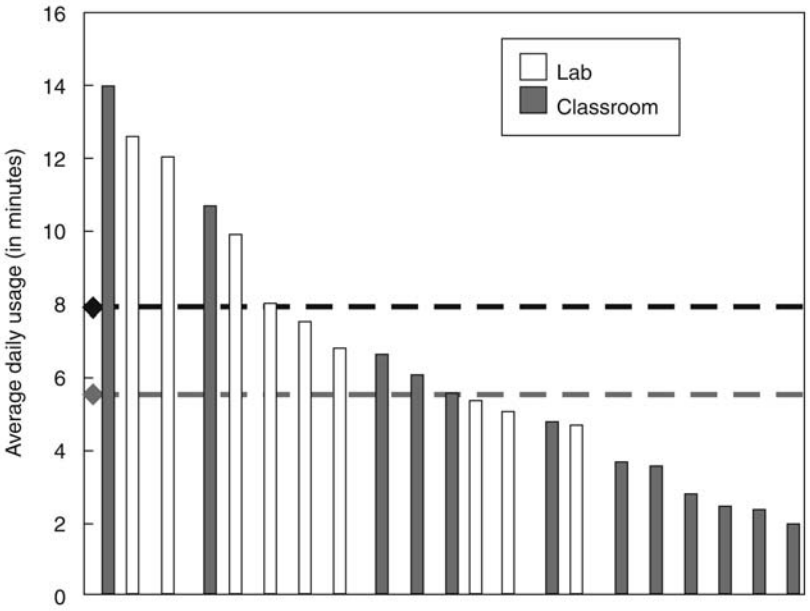

Figure 7.3. Per-Student Usage in 2002–2003, by Room

Table 7.3. Effects of Context on (Unadjusted) Usage

Fraction of Variance Explained	*Frequency*	*Duration*
Setting	.154	.171
School	.547	.342
School + Setting	.562	.348
Teacher	.809	.694

Note: All effects shown are statistically significant.

the labs, versus only 5.32 ± 3.66 minutes per day for the classrooms. Third, usage varied dramatically between classrooms. At the low end, 6 of the 12 classrooms averaged below 4 minutes per day. At the high end, the room with the very highest usage (13.9 minutes per day) was a classroom setting, not a lab. We conclude that most teachers found whole-class lab implementations easier than classroom implementations, but the teachers most committed to using the Reading Tutor managed to achieve high usage in their classrooms.

Table 7.3 compares the effects of *setting*, *school*, and *teacher* on usage in terms of the fraction of variance they account for. Overall, *setting* is only a moderate predictor of usage, accounting by itself for less than 20 percent of variance in session frequency or duration, and adding virtually no variance beyond what is explained by school. Apparently *school* is much more important than *setting*. In contrast, *teacher* is a very important influence, explaining about 25 percent additional variance in session frequency and about 35 percent in session duration. However, *teacher* is hard to disambiguate from *school* because the data include so few (two to three) teachers per school. *Grade* had no significant overall effect on usage, and student ability had a negligible effect, accounting for only 2 percent of variance.

Table 7.4 uses correlations to quantify the influence of grade, ability, and student-computer ratio within classroom and lab settings. Session frequency and duration both increased significantly with grade in classroom settings but decreased with grade in lab settings. Ability correlated negatively with session frequency in classrooms but positively with increased session duration in labs. It is hard to explain these findings other than as artifacts of the particular teachers involved.

We correlated the number of students per Reading Tutor against both measures of usage. In the classroom setting, this ratio had significant and substantial negative correlations with both usage measures, indicating that student-computer ratio matters in classrooms. In the whole-class lab setting, the correlation was negligible for session duration and *positive* for session frequency, presumably indicating that student-computer ratio does not matter in a lab so long as there are more computers than students.

Table 7.4. Correlations within Setting of Usage with Grade, Ability, and Number of Students per Computer

	Frequency	*Duration*
Grade		
Class	**.21**	**.471**
Lab	**−.20**	**−.241**
Ability		
Class	**−.430**	−.123
Lab	.038	**.243**
Ratio		
Class	**−.356**	**−.384**
Lab	**.267**	.095

Note: Bold correlations are significant at $p < .01$.

Analyzing Efficacy

Although the Reading Tutor is effective, evidence suggests that it is especially effective for some students (Mostow, Aist, and Burkhead et al. 2003; Mostow, Aist, and Huang et al. forthcoming; Mostow and Beck, forthcoming). This result prompts three questions: (1) For which students is the Reading Tutor most efficacious? (2) How can we scale the Reading Tutor's efficacy across a broad range of students? That is, which Reading Tutor behavior contributes to its efficacy for which students? and (3) How does student behavior affect Reading Tutor efficacy? By *efficacy*, we mean the gain achieved by a specified amount of Reading Tutor use.

Speech-recognition-based, computer-guided oral reading has demonstrated usability, user acceptance, assistive effectiveness, and even pre- to posttest gains (Aist and Mostow 1997c; Cole et al. 1999; Mostow, Roth, and Hauptmann et al. 1994; Nix, Fairweather, and Adams 1998; Russell et al. 1996; Williams 2002; Williams, Nix, and Fairweather 2000), but the proof of the pudding is whether it significantly *increases* learning gains over gains that children make otherwise. Even with barely twenty minutes of use per day, successive versions of the Reading Tutor have produced substantially higher comprehension gains than current practices in controlled studies lasting several months. We use valid and reliable instruments to measure gains from pretest to posttest. We record the following student variables: age, grade, phonemic awareness (Wagner, Torgesen, and Rashotte 1999), reading skills (Torgesen, Wagner, and Rashotte 1999; Wiederhold and Bryant 1992; Woodcock 1998), spelling ability (Larsen, Hammill, and Moats 1999), motivation (McKenna, Kear, and Ellsworth 1995), and special needs status.

Which Students Benefit Most?

These studies enable us to determine which students benefit most from the Reading Tutor. Study designs at each school depend on time, space, and policy constraints that affect whether Reading Tutors are deployed in classrooms, labs, or specialists' rooms; which students participate in the study; and which current practices are acceptable as alternative treatments. To control for teacher effects, we use within-class study designs wherever possible, so as to ensure that results are due to the Reading Tutor intervention. To control for student differences when comparing different treatments within the same classrooms, we assign students randomly, stratified by pretest scores, to either the Reading Tutor or an alternative treatment.

When necessary we conduct between-class designs. One benefit of comparing between-class designs versus comparing within-room designs is ecological validity between class designs; they avoid artifacts specific to having some of the students in a class use the Reading Tutor but not others. In particular, between-class designs account for the Reading Tutor's impact on classroom instruction. If teachers learn from the Reading Tutor's reports that several students have a specific deficit in reading skills, they may adjust their instruction to remediate it—thereby benefiting other students who have the same problem, including students who do not use the Reading Tutor. More simply, the Reading Tutor may free up the teacher to give more individual attention to students who do *not* use it. Within-class comparisons would not detect such effects.

We compute effect size as the *difference* in gains between the Reading Tutor and current practice, divided by the average standard deviation in gains of the two groups. Effect sizes for passage comprehension are substantial compared to other studies (NRP 2000): 0.60 for sixty-three students in grades two, four, and five at a low-income urban school (Mostow and Aist 2001; Mostow, Aist, and Huang et al. forthcoming); 0.48 for sixty-six third graders at a lower-middle-class urban school (Aist, Mostow, and Tobin et al. 2001; Mostow, Aist, and Burkhead et al. 2001; Mostow, Aist, and Burkhead et al. 2003); and 0.66 for fifty-two first graders at two suburban Blue Ribbon Schools of Excellence (Mostow, Aist, and Bey et al. 2002; Mostow and Beck, forthcoming). The cited publications report additional results.

It is important to point out that these studies analyze overall effectiveness, which depends on usage as well as efficacy. Teasing apart efficacy from usage is problematic. We cannot simply compute efficacy by correlating gains against usage, because usage is not a random variable. Rather, usage is influenced by many variables that also affect gains, such as students' attitude and attendance, and teachers' classroom management skills. As of yet, we have not attempted to manipulate the amount of usage as an experimental

variable in order to measure dosage effects. We are still grappling with how to achieve a given target level of usage, which is quite hard enough.

Which Reading Tutor Actions Contribute to Efficacy?

To analyze which tutorial actions help which students in which cases, we use continuous assessments of student progress and automated embedded experiments within the Reading Tutor. Automated experiments embedded in the Reading Tutor evaluate its own interventions in a way that combines the methodological rigor of controlled experiments, the ecological validity of school settings, and the statistical power of large samples. The network of Reading Tutors automatically administers thousands of randomized trials to test the effects of tutorial actions on student performance (Aist 2001b; Mostow, Tobin, and Cuneo et al. 2002c), records their outcomes, transmits them to our lab, and aggregates them for subsequent analysis (Mostow, forthcoming). The key idea is to check for the signature of tutorial actions on the student's performance. To compare the effects of different tutorial actions (including none) on a given skill, an embedded experiment randomly chooses which action to perform. Aggregating over many such trials lets us compare student performance outcomes and discover not only which tutorial actions work best overall, but under what conditions (Aist 2002a, 2000b; Aist and Mostow 1998; Mostow and Aist 2001). Matched trials amplify the power of this method. For example, an experiment that compared five different methods for previewing new words in a story randomly assigned each method to a different word in the story, thereby matching on both student and story (Mostow, forthcoming).

Such experiments measure assistive effects of scaffolding by analyzing the student's immediate performance, and trace longer-term learning effects by analyzing the next opportunity to demonstrate the skill (Corbett and Anderson 1995). This opportunity can be explicitly scheduled as a delayed posttest, occurring from minutes to days later, or defined naturalistically, for example, based on how the student performs on the next encounter with the same word in a different story, or with a similar word. Other researchers have used explicit tests of recently taught words, older taught words, and similar untaught words, to help evaluate decoding instruction (Sharp, Goldman, and Bransford 2002). By parsing longitudinal tutor-student interactions into successive encounters of each word or other unit of skill, we get millions of data points to measure the assistive and learning effects of tutorial actions and, as IERI (2002) asks, "test, in actual school settings, the validity of newly discovered knowledge of important aspects of reading," including phonemic awareness, decoding, vocabulary, fluency, and comprehension.

For example, explicit vocabulary instruction is important but time consuming (Beck, McKeown, and Kucan 2002). Explaining unfamiliar words and concepts in context can remedy deficits in vocabulary and background knowledge (Elley 1989) by scaffolding the reader with information at the "teachable moments" when it is needed. Accordingly, we tried supporting vocabulary acquisition by presenting short "factoids"—comparisons to other words (Aist 2001b, 2002a). An automated experiment embedded in the 1999–2000 Reading Tutor tested the effectiveness of reading a factoid just before a new word in a story, compared to simply encountering the word in context without a factoid. The outcome variable was performance on a multiple-choice question, presented the next day the student used the Reading Tutor. Analysis of over 3,000 randomized trials showed that factoids helped on rare, single-sense words, and that they helped third graders more than second graders (Aist 2000a, 2001a, 2001b). A follow-on experiment (Mostow, Beck, and Bey et al. 2003; Mostow, Beck, and Bey et al. 2004) showed that presenting synonyms or short definitions of new vocabulary before a story improved performance both on answering multiple-choice closed questions within the story, and on matching words to their definitions after the story—but only for students above a certain level of reading proficiency. More generally, by acquiring predictive models of the effects of tutorial actions, embedded experiments can inform a decision-theoretic approach to tutoring (Beck 2001, 2002; Beck and Woolf 2000, 2001; Beck, Woolf, and Beal 2000; Murray, VanLehn, and Mostow forthcoming).

Which Student Behaviors Affect Efficacy?

The effects of the student's behavior on gains are harder to analyze than the Reading Tutor's, because we cannot directly control student behavior in order to conduct randomized experiments. However, we can still use correlational analyses.

One way to analyze students' behavior is to examine how the students allocate time among different types of actions in the Reading Tutor: logging in, picking stories, reading, writing, waiting for the Reading Tutor to respond, and so forth. Mostow, Aist, and Beck et al. (2002) partial-correlated students' pre- to posttest gains against the percentage of time they spent on each such action, controlling for pretest score differences among students so as to exclude prior proficiency as a confounding variable.

Analysis of fluency gains in a 2000–2001 controlled study found a +0.42 partial correlation with percentage of time spent actually reading, and a -0.45 partial correlation with percentage of time spent picking stories (both statistically significant). These results do not establish conclusively that the observed behavioral differences caused the differences in gains. However,

the outcome differences were not predicted by any of the pretests, which included a measure of attitude toward reading (McKenna, Kear, and Ellsworth 1995).

CONCLUSIONS

The takeaway message of this chapter can be summarized by the formula "Effectiveness = Usage × Efficacy," coupled with the use of technology to instrument all three variables. The Reading Tutor's effectiveness is the increase in gains it produces compared to what it replaces. Its usage is the product of session frequency and session duration, both of which depend on contextual influences in ways we can quantify based on data it captures. Its efficacy is influenced by behavior, both its own and students', in ways we can analyze using experimental and correlational methods, respectively. We hope that the practical lessons and research methods discussed in this chapter will prove useful to others as well.

NOTE

1. The title of this chapter is borrowed from a joint invited talk with Dr. Gregory Aist at the "Workshop on Bridging the Digital Divide for Work and Play," held November 3–4, 2001, in Toronto, Ontario, Canada. However, the material here is new. We thank the students and educators who participated, and current and past members of Project LISTEN who contributed, especially assistant director Roy Taylor, D. Ed., for negotiating and documenting implementations; Educational Research field coordinator Kristen Bagwell, and field support specialist Julie Sleasman for supporting sites and supervising pre- and posttesting; research programmer Andrew Cuneo for developing the Reading Tutor and its database; Amy Quinn for implementing the teacher report tool, and the team of Human-Computer Interaction students who helped design it (Alpern et al. 2001); Joseph Valeri and Juliet Bey for developing the Reading Tutor's interactive materials and implementing its automated experiments; and University of Pittsburgh professor Rollanda O'Connor for her expertise on reading and contributions to the IERI proposal on which some passages in this chapter are based. This work was supported in part by the National Science Foundation under Grants REC-9979894 and REC-0326153. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation or the official policies, either expressed or implied, of the sponsors or of the United States government.