Experimentally Augmenting an Intelligent Tutoring System with Human-Supplied Capabilities: Adding Human-Provided Emotional Scaffolding to an Automated Reading Tutor that Listens

Gregory Aist[1], Barry Kort[2], Rob Reilly[2], Jack Mostow[3], and Rosalind Picard[2]

[1] With Project LISTEN 1996-2001; now Visiting Scientist, MIT Media Lab;
and Research Scientist, Research Institute for Advanced Computer Science,
Mail Stop T27-A, NASA Ames Research Center,
Moffett Field CA 94035-1000 USA
GregoryAist@yahoo.com

[2] MIT Media Lab, Building E15, 77 Massachusetts Ave., Cambridge MA 02139 USA
http://affect.media.mit.edu/AC_research/lc/

[3] Project LISTEN, RI-NSH 4213, Carnegie Mellon University,
5000 Forbes Ave., Pittsburgh PA 15213
http://www.cs.cmu.edu/~listen

Abstract. This paper presents the first statistically reliable empirical evidence from a controlled study for the effect of human-provided emotional scaffolding on student persistence in an intelligent tutoring system. We describe an experiment that added human-provided emotional scaffolding to an automated Reading Tutor that listens, and discuss the methodology we developed to conduct this experiment. Each student participated in one (experimental) session with emotional scaffolding, and in one (control) session without emotional scaffolding, counterbalanced by order of session. Each session was divided into several portions. After each portion of the session was completed, the Reading Tutor gave the student a choice: continue, or quit. We measured persistence as the number of portions the student completed. Human-provided emotional scaffolding added to the automated Reading Tutor resulted in increased student persistence, compared to the Reading Tutor alone. Increased persistence means increased time on task, which ought lead to improved learning. If these results for reading turn out to hold for other domains too, the implication for intelligent tutoring systems is that they should respond with not just cognitive support – but emotional scaffolding as well. Furthermore, the general technique of adding human-supplied capabilities to an existing intelligent tutoring system should prove useful for studying other ITSs too.
Keywords: emotional scaffolding, affective computing, guided oral reading, intelligent tutoring systems.

1 Introduction

Educators have long been aware of the importance of motivational factors in learning. Beyond such intuitive awareness, Coles (1999) notes that "Some connections between learning and emotions have been well documented: for example, poor learning can produce negative emotions; negative emotions can impair learning; and positive emotions can contribute to learning achievement and vice versa" (Coles 1999). Neuroscience and cognitive science have made important strides in recent years unraveling the mystery of the role of emotions in cognition, decisions, and learning (e.g., Damasio 1994, Isen 2000).

Today, researchers are mapping out the emotional factors that arise in the course of learning, in an effort to craft more effective educational resources which stimulate and maintain the motivation to learn (Lepper et al. 1993; Kort et al. 2001). Lepper et al. (1993) analyzed how expert tutors use motivational strategies; del Soldato and Boulay (1995) proposed incorporating motivational strategies into tutoring systems. In this paper, we look at adding human-supplied emotional scaffolding to an existing intelligent tutoring system – as a means of exploring the potential benefits of enabling intelligent tutoring systems to be emotionally appropriate.

But what does it mean for an intelligent system to be emotionally appropriate? We can consider first the capability of *sensing* affective information, and separately the capability of *responding* in emotionally appropriate ways. A system that can *sense* affective information (either passively, e.g. by computer vision, or actively, e.g. by asking) may or may not reveal its awareness much of the time. A system that can *respond* in emotionally directed ways makes decisions based on perceptions and prior knowledge about how best to respond or to not respond, taking into account the student's affective state. This problem of deciding how to respond is a qualitatively different research problem from the sensing problem. Thus we can place emotionally appropriate computational systems on the chart shown in Table 1.

	Does not sense/perceive affective information	Senses/perceives affective information
Does not supply emotionally directed responses	The norm for computer software.	Example: System silently records examples of user frustration, perhaps for later review by quality assurance team.
Supplies emotionally directed responses	Example: System supplies spoken apology triggered by network delay	Example: System detects student frustration; presents a slightly easier task.

Table 1. Degrees of kindness: affective computing in sensing and acting.

Ultimately, affective computing systems should both sense and respond to emotions. Research into automated sensing of emotions is underway as described elsewhere (Picard 1997). In the meantime, we would like to further the research agenda of automated responding without having to wait for automated human-level sensing of emotions – and to take advantage of whatever opportunities for improving intelligent tutoring systems may present themselves along the way. A promising way to start is to identify specific aspects of state-of-the-art intelligent tutoring systems where emotional scaffolding (provided at least initially by human wizards) can substantially improve over current practice. In this paper, we present one such study exploring emotional awareness in ITSs, carried out as a cooperative effort between Carnegie Mellon's Project LISTEN (Mostow & Aist FF 2001) and MIT's The Affective Learning Companion (Picard 1997; Kort et al. 2001).

The remainder of the paper is as follows. First, we describe a series of quick, informal pilot user studies that we conducted in order to verify underlying assumptions, evolve our research questions, and refine the experimental protocols. Second, we describe the main study we carried out, and its principal results. We then discuss further implications for intelligent tutoring research.

2 Computer Tutor plus Human Wizard

Our fundamental research question for this paper was "In what ways can human-level emotional scaffolding improve on the state of the art in intelligent tutoring systems?" In order to address this question, we used techniques drawn from ablation

studies and Wizard of Oz studies. Ablation studies, a common method for studying intelligent tutoring systems, take a working system and deconstruct it, systematically disabling various components and examining the effectiveness of the remaining system. The Wizard of Oz method, a familiar technique for designing interactive systems, replaces the "brains" of a yet-to-be-built computer program with a human wizard, pulling the strings from behind the scenes. In some variations, the wizard automates him- or herself out of the loop (wholly or partially).

Our experiments were as follows. First, we inverted the ablation study. We focused on adding components and testing for increased effectiveness, rather than removing components and testing for decreased effectiveness. Second, we did not want to embed one particular hypothesis in stone (that is, C++ code). Therefore, we imbued Project LISTEN's Reading Tutor with a human wizard filling in the emotional gaps. We thus created a hybrid with a human in the loop: Reading Tutor Plus.

(We note that this methodology owes a debt of inspiration in part to some sections of Alex Rudnicky's proposal to develop automated techniques to allow a computer to gradually learn how to do computer technical support (Rudnicky, internal presentation to CMU Sphinx group, fall 2001).)

In this study in particular, we began with Project LISTEN's Reading Tutor, which listens to children read aloud, and helps them learn to read (Mostow & Aist FF 2001; Figure 1). The Reading Tutor helps a student read a story as follows. The Reading Tutor displays a story one sentence at a time, and listens to the student read all or part of the sentence aloud. The student can click *Back* to move back to a previous sentence, and *Go* to move on. The Reading Tutor responds with help modeled in part after expert reading tutors, but adapted to the strengths and weaknesses of computer technology. The Reading Tutor provides praise for good or improved reading.

In our experiments, the affective component consists of the wizard injecting emotionally directed remarks over the same audio circuit by which the student hears the Reading Tutor – without, so far, controlling the Reading Tutor's behavior directly as is normally done in Wizard of Oz studies. In the initial study design, both the Reading Tutor's own praise and the human-supplied emotional scaffolding were active. For reasons we describe later, in the main study design, the Reading Tutor tutored the same as it ordinarily would – but without the praise components previously introduced into the system. But what should the human wizard say? If the human wizard can say anything at all, the only reasonable conclusion from statistically reliable differences between trials with human support and those without (we argue) is that the difference is due to the (active) presence of the human wizard. While probably true, such a distinction is not terribly interesting since it does not shed any light on exactly *what* the wizard did that was helpful; nor is that distinction terribly useful since it does not shed any light on exactly *how* to automate the wizard's scaffolding.

We therefore set about to identify concrete ways in which to limit what the wizard said, while still yielding a protocol that we expected to have significant and substantial impact on the tutoring process.

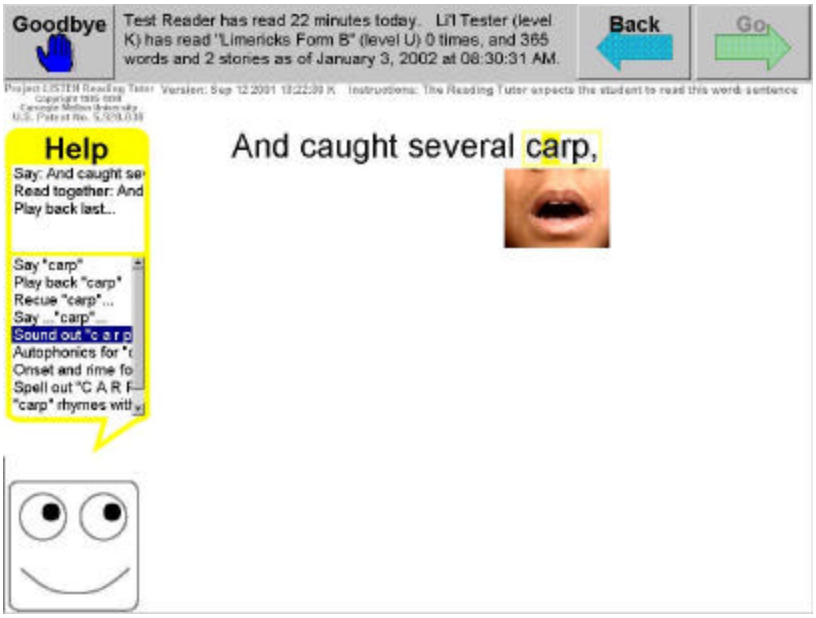

Figure 1. Screenshot from Project LISTEN's Reading Tutor. The Reading Tutor is in the middle of sounding out *carp*.

3 Pilot Studies

In order to acquire a set of age-appropriate, relatively similar texts, we decided to adapt a set of limericks written in the 19th century by Edward Lear for research use (see Aist (2000) for previous use of these limericks.) Each limerick contained a single difficult word. The first question we decided to investigate, in late August 2001, was whether an encounter with a new and difficult word would trigger a visible emotional response. We used a within-subject design consisting of repeated trials with words in different conditions for different subjects. This methodology, familiar in medical research as a repeated trials design, has been adapted to intelligent tutoring research in the form of experiments embedded into the course of normal student-Reading Tutor interaction (Mostow and Aist FF 2001; Mostow et al. NAACL 2001). Students read eight children's limericks, four in each of two conditions, as follows:

(experimental) the limerick contained the original difficult word, such as *irascible*, or
(control) the limerick contained a substituted easier word or phrase, such as *grouchy*.

The conditions were designed so that the experimental condition would elicit a visible emotional response (such as frustration), but the control condition would not. We collected videotaped data from five students. (The first two students' data is more or less unusable due to bad or missing audio.) We wanted to see whether *any* experiment seeking to manipulate difficulty in order to trigger visible student responses (at the level of individual words) might succeed. Thus for this first pilot study we simply watched the videotapes and formed informal impressions. (Our follow-on study, as reported later in this paper, used more systematic experimental techniques.) Our impressions of the pilot data were as follows: Students did indeed occasionally respond visibly to difficult words, but infrequently; furthermore, whether a response was forthcoming did not seem to be uniformly predictable whether for a given student or a given word. Nonetheless, we were relieved that these observations bolstered one of the foundational assumptions of embedding emotional awareness into intelligent tutoring systems: difficult tasks encountered on an intelligent tutoring system can sometimes trigger visible emotional responses.

We then sought to develop a protocol whereby the human wizard could intervene at specific limited points with emotionally appropriate responses, and the outcome measures would include measures of learning.

In late September 2001, we set up a protocol where students read a series of limericks interspersed with definitions of the hard (target) word. For example:

There was an Old Man of Cape Horn,
Who wished he had never been born;
So he sat on a chair,
Till he died of despair,
That dolorous Man of Cape Horn.
We can say a person is dolorous if they are mournful, or feel really bad.

The human wizard could see and hear the student over a dedicated television link, and was initially restricted to providing emotional scaffolding, such as praise or encouragement, between the limerick and the definition. The student could see and hear the Reading Tutor, and could hear but not see the wizard. The control condition lacked such emotional scaffolding. Both conditions included the previously developed Reading Tutor-supplied praise as described in Mostow and Aist (CALICO 1999) and subsequently updated in later versions of the Reading Tutor.

Our experience after working with just a few students reading under this protocol, however, showed clearly (at least to us) that this limited amount of emotional scaffolding was simply not enough to reasonably expect any difference between the emotionally supported (experimental) trials and the neutral (control) trials – no difference in moment-by-moment observable behavior would, we suspect, mean no differences in student behavior over the course of an entire session. Furthermore, the presence of Reading Tutor-supplied praise might very well confound any effects of human-supplied praise.

4 Revised Protocol

Therefore, we designed a revised protocol. We decided to greatly expand the times when emotional scaffolding was permitted, as follows. First, we made the points when the wizard could provide emotional scaffolding include not just between limericks and definitions, but between *every* pair of sentences. We also allowed the wizard to intervene to alleviate frustration with a difficult sentence by suggesting that the student move on to the next sentence: a "mercy rule" to maintain a good pace of reading. Finally, we suppressed the Reading Tutor's built-in praise, so that such praise would not interfere with the wizard in the experimental condition – or dull the contrast between Reading Tutor plus wizard vs. Reading Tutor alone. Why not include the unablated Reading Tutor as an alternate control condition? The number of students we were able to run in this study was too small to permit three conditions; we decided that the comparison of most interest was the Reading Tutor with no emotional scaffolding vs. the Reading Tutor with human-supplied emotional scaffolding.

We used two sets of 10 limericks, with each set sorted in approximate order of the difficulty of the target word. In previous work we had found that second and third graders did not remember much from these limericks when tested on a later day (Aist 2000.) This suspicion that second and third graders were too young for these limericks was confirmed when we tried to have one second grade student read them; we agreed that she was in over her head.

Thus we decided to come up with a different text for the second and third graders: We had these students read a list of progressively harder words.

We decided to get a list of words that would progressively stretch children's decoding skills. Thus we constructed the list(s) of progressively harder words (Form A, and Form B) as follows.

First, we selected those words with a frequency of two per million words, as given by the Brown corpus (Kucera and Francis 1967).

Second, we excluded those words marked by the MRC Psycholinguistic Database as usually capitalized, for example *Lionel*, *Belgium*, *Unitarianism*.

Third, we sorted the remaining words by the number of syllables and the number of letters.

Fourth, we selected by hand several sets of words, preferentially selecting decodable words (those which can be read left-to-right according to common letter-to-sound rules of English.) The words ranged from one-syllable, three letter words like *bib* and *den* to six-syllable, thirteen- and fourteen-letter words like *acceptability* and *transcendentalism*.

Fifth, we paired the words (e.g. *complimentary* paired with *differentiate*; *acceptability* paired with *accessibility.*)

Sixth, we randomly assigned one word of each pair to Form A, and the other word of each pair to Form B. Each word on Form A thus had a twin on Form B.

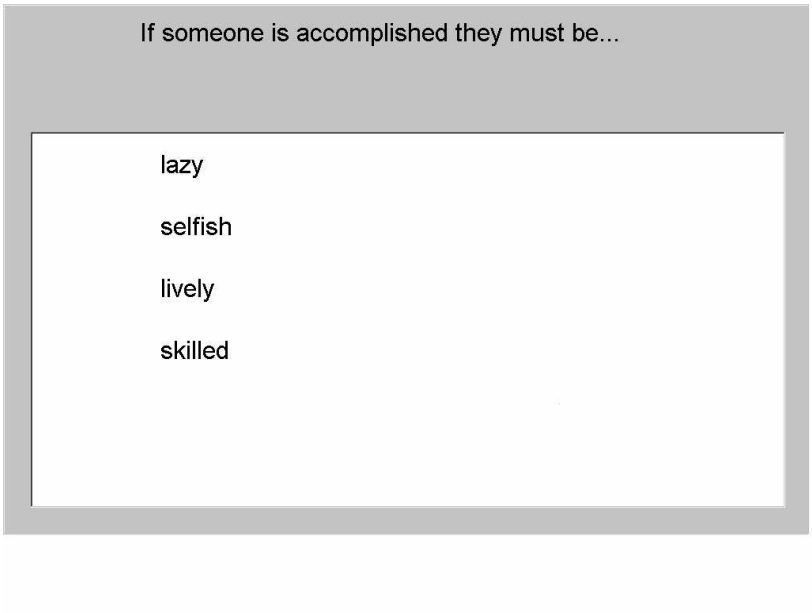

Figure 2. Vocabulary question after each limerick.

One of our expectations was that emotional scaffolding would lead to increased student attentiveness, which would lead to increased retention of the vocabulary information presented in the limericks and accompanying vocabulary explanations. Another expectation was that human-provided emotional scaffolding would increase students' persistence. Normally, the length of time that a student uses the Reading Tutor is strongly affected by the length of the story, the length of the class period, and other factors essentially out of our control.

In order to provide for an explicit measure of memory of the word that it had just explained, the Reading Tutor asked a vocabulary question after each limerick (Figure 2.) There were no questions during the wordlist task. In order to provide for an explicit measure of persistence, we inserted choice points into the wordlist and limerick tasks. After every ten words (for the wordlists) or after every limerick, including its corresponding definition and vocabulary question, (for the limericks), the Reading Tutor displayed as text and spoke aloud a prompt inviting the student to click Goodbye to stop, or Go to continue reading. It then alternately flashed the Goodbye and Go buttons six times.

We determined who would receive which condition when as follows. In each study, we matched students by classroom (and thus grade), and by gender. Each student read one of the versions on (their own) Day 1, either with or without emotional scaffolding. Approximately one week later, that student read the other version in the other condition. We counterbalanced treatment by version and order.

Figure 3. Experimental and control conditions for computer tutor plus human wizard experiment. A mirror in front of the student allows the video camera to capture the laptop monitor and the student's face simultaneously.

We measured the following dependent variables.

- *persistence*: the number of times the student clicked Go to continue reading instead of clicking Goodbye to quit. We would expect persistence to be important because increased persistence means more time on task – a key ingredient of success in reading.
- *memory*: the number of words a student got right on limericks quiz.

Figure 3 shows the experimental and control conditions: in the experimental condition, a human wizard observed the student on a television monitor in an adjacent room and provided emotional support over the student's headphones. In the control condition, the wizard helped the student log in and choose the correct story – but after that simply observed, and the Reading Tutor provided help as usual (minus its pre-existing emotional support, which had been suppressed for this experiment.)

The human wizard provided emotional support as follows.

1. Selecting phrases from a web page, to play pre-recorded audio.
2. Speaking live into a microphone to provide help that was (not yet) available in pre-recorded form.

As the experiment progressed, the wizard added phrases used in the previous days' experiment to the web page – automating himself (partially) out of the loop.

To characterize the interaction between the human wizard, the student, and the Reading Tutor, we transcribe below an extended example of one student using the Reading Tutor without emotional support, and the same student using the Reading Tutor while receiving emotional scaffolding from the human wizard. We transcribed these sessions on a turn-by-turn basis from videotape, noted the begin and end times for the first set of words, and added a brief description of the student's behavior when faced with the choice to continue or quit.

Student[1] without emotional support October 31, 2001 Wordlist Form A	Same student with emotional support October 17, 2001 Wordlist Form B
(Wizard directs student to choose "Word List Form A".) RT displays text "Word List Form A" RT says "Word" W: "Can you read that?" S: "Word" RT: "Word" S: "List Form A" RT displays and speaks: *Now you'll see some words to read*. S clicks Go RT displays and speaks: *Please read each word aloud.* S: "Please read each word aloud." S clicks Go. 11:49:14 AM RT displays: *den* S: "den" S clicks Go RT displays: *eel* RT says "starts like *eels*" S: "ee- eel" RT displays: *fry* RT says: rhymes with *wry* S: "fry?" RT displays: *gym* RT sounds out *gym* S: "gym" S (again): "gym" S (yet again): "gym" RT displays: *kin* RT: n here makes the sound /n/ (At this point the RT may have heard itself, and thus went on to the next word right away.) RT displays: *kit* RT says: "rhymes with lit" S: "kit" RT displays: *pal* RT sounds out: "al" S: "pale" RT displays *ram* RT sounds out *ram* S: "ram" RT displays *spa* RT breaks *spa* down into onset and rime S: "spa" RT: displays *vow* S: "vow" RT displays and says: You can click Goodbye to log off, or Go to continue reading." (Goodbye and Go alternately flash.) 11:50:48 AM S narrows eyes, scratches chin, then clicks Goodbye.	(Wizard directs student to choose "Word List Form B".) RT displays text "Word List Form B" RT says "Word List Form B" RT says "Starts like World" RT says "Word List Form B" **W: "Can you read that?"** **W: "Can you read that out loud, Danny?"** RT: sounds out *form* S (overlapping): "Word List From B" [sic] **W: "Good."** RT displays and speaks: *Now you'll see some words to read*. **W: "You can read this aloud, or click Go."** S: "Now you'll see some words to read." **W: "Very good."** S clicks Go. RT displays and speaks: *Please read each word aloud*. S: "Please read each word aloud." **W: "Good."** 11:53:27 AM RT displays: *bib* S says: "bib" S clicks Go RT displays: *fin* RT sounds out *fin* S says: "fin" S clicks Go RT displays: *fad* RT says: "Rhymes with *fats*" S: "fad" **W: "You're doing fine."** RT displays *gel* RT sounds out *gel* S: "gel" **W: "Good."** RT displays *hop* S: "hop" S clicks on Go. RT says: "Sorry, can't go on right now." RT displays: *lug* S says "jug" RT: mmm RT says "lug" S: "lug" **W: "Good."** RT displays: *maw* RT says "a w here makes the sound aw" S: "w – maw" **W: "Good."** RT says "maw. maw." S: "maw" RT displays: *pup* RT says "starts like *pun*" RT: mmm

[1] Subject label: mDS.

	S: "pawp" RT displays *sip* RT breaks *sip* down into onset and rime **W: "Well done."** RT displays: wed RT sounds out wed S: "wed" RT displays and says "You can click Goodbye to log off, or Go to continue reading." (Goodbye and Go alternately flash.) **W: "You're reading very well."** S's eyes go back and forth between Go and Goodbye RT: "You can c-" 11:55:15 AM S clicks Go. (and thus goes on to read more words.)

5 Results

We present the results from the outcome variables *persistence* and *memory*. Out of 17 students who began the study, 14 students completed it, as shown in Table 2.

Table 2. Persistence and memory results from study comparing Reading Tutor alone to Reading Tutor plus human-supplied emotional scaffolding.

Number	Gender	Persistence (Control)	Persistence (Experimental)	Correct (Control)	Correct (Experimental)	Word List / Limericks	Which story first?
1	F	15	15			Word List	Control
2	F	10	10	0.6	0.8	Limericks	Control
3	F	10	9	0.6	0.44	Limericks	Experimental
4	F	.	1			Word List	Experimental
5	F	10	10	0.4	0.6	Limericks	Experimental
6	F	16	16			Word List	Experimental
7	F	5	2	0.4	0	Limericks	Control
8	F	4	15			Word List	Control
9	M	6		0.83		Limericks	Control
10	M	16				Word List	Control
11	M	6	8	0.83	0.5	Limericks	Experimental
12	M	1	5			Word List	Experimental
13	M	3	10	0.33	0.6	Limericks	Experimental
14	M	1	6	1	0.83	Limericks	Control
15	M	6	10	0.67	0.9	Limericks	Control
16	M	1	2			Word List	Control
17	M	1	2			Word List	Experimental

Figure 4 shows *persistence* in the control condition vs. *persistence* in the experimental condition – each point represents one student except where noted. *Persistence* in the control condition averaged 6.4 ± s.d. 5.1 for wordlists and limericks combined; in the experimental condition, 8.6 ± s.d. 4.8 for wordlists and limericks combined, a significant difference ($p = 0.041$ by 2-tailed paired T test; $F = 5.07$, $p = 0.044$ by repeated measures analysis of variance with story kind (wordlist or limericks) as a between-subjects factor; $F = 4.46$, $p = 0.061$ by repeated measures analysis of variance with story kind and order as between-subjects factors) with an effect size of 0.44.

Boys persisted on average twice as long with emotional scaffolding as without; however, many of the girls were already at or near maximum persistence:

Boys: N=7, 6.1 ± 3.4 (expt.) vs. 2.7 ± 2.4 (control), p=.007 by paired T -test.

Girls: N=7, 11.0 ± 4.9 (expt.) vs. 10.0 ± 4.5 (control), p=.582 by paired T-test.

No significant differences were found on *memory* (Figure 5), which was measured only for the limericks.

One might suggest that it wasn't the emotional support that made the difference, but rather that the student knew a real human was present and aware of what the student did. However, in this experiment the wizard was in fact present during the control condition, but served merely as an observer. The difference between conditions therefore cannot lie in presence alone, but rather somewhere in the realm of supplying emotional scaffolding.

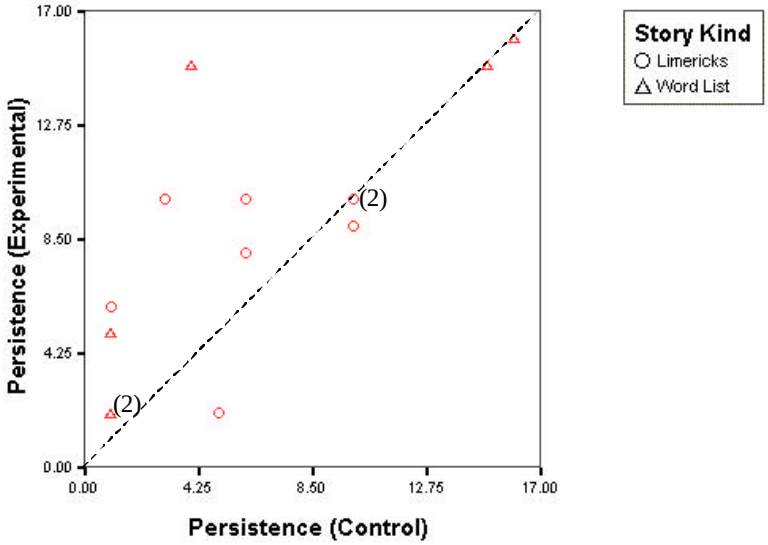

Figure 4. Persistence in the control condition and in the experimental condition.

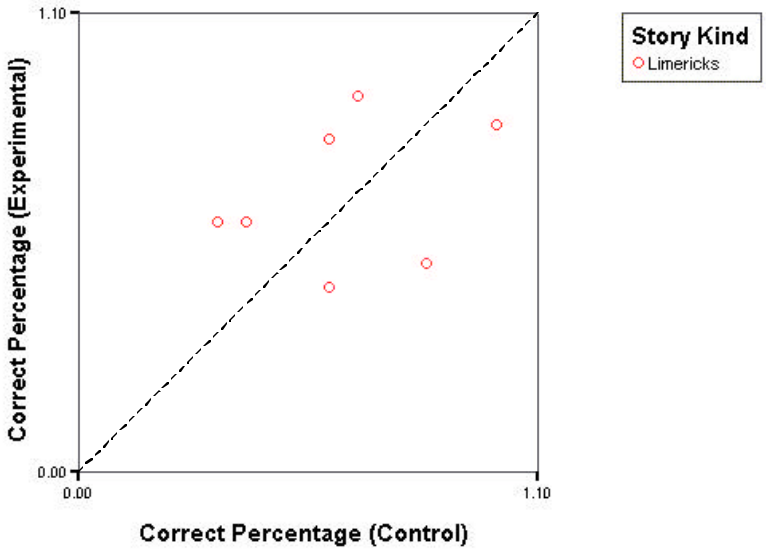

Figure 5. Memory in the control and experimental conditions (limericks only).

We had initially noted some signs that emotional support might lead to prolonged engagement. As the wizard said later:

> At least a few subjects seemed to be responsive to emotional scaffolding. There were a few subjects in the Experimental Condition who seemed to remain engaged longer as a result of judiciously timed encouragement just as they were deciding whether to continue or quit. (Barry Kort, via intragroup email, November 9, 2001).

However, as the experiment progressed, we had reason to suspect that we were going to end up with student frustration swamping any other effects of affect. As the wizard later recalled:

> The most frequent source of frustration was when the RT failed to accept a perfectly articulated sentence, phrase, or word. There were selected words that it consistently failed to recognize time and again. Some subjects repeated the fragments with increasing levels of annoyance. Some even cursed at the machine. The other annoyance was the long delays before it advanced to the next fragment. (Barry Kort, via intragroup email, November 9, 2001).

Thus it was gratifying to see that emotional scaffolding did in fact lead to significantly greater persistence (for boys). In fairness to the Reading Tutor, we should point out that mitigating factors include laptop sound card problems, and performance slowdowns caused by memory leaks and full disk. Furthermore, these problems – and their severity – don't seem representative of the Reading Tutor's normal performance in the field. We do see false alarms and frustration. However, they're mitigated by technical factors (the desktop computers used for 84 of the 99 Reading Tutors in daily use at schools for the 2001-2002 school year don't have many of the problems we found on the laptop used in this experiment, some of which have now been fixed), by context (the students get to take turns picking what to read – not just limericks or word lists), and by the Reading Tutor's built-in emotional support (the unablated Reading Tutor praises students (Mostow & Aist CALICO 1999)).

One kind of result from this study was simply the types of emotional support produced – fodder for possible later automation. The wizard produced various kinds of support. For example, the wizard sometimes provided praise such as "Good" or "Very good", or encouragement such as "You're doing fine." For example, after the system offered a choice to click Go to continue, or Goodbye to log off – when the student was deciding whether to continue reading or not, the wizard sometimes would supply judiciously timed encouragement such as "You're doing fine." This and other kinds of encouragement ought to provide a rich source of potential Reading Tutor interventions.

It was also interesting to note the interactions between normal Reading Tutor behavior and the interventions supplied by the wizard. For example, when words in the wordlist became excessively difficult, the Reading Tutor modeled the pronunciation so that the student didn't even have to read the word. In the word-list exercise, the RT typically spoke every word upon display after the first few sets of words. Many subjects stopped looking at the screen and just repeated what they heard. The Reading Tutor normally provides assistance both preemptively and on demand – but in this case this assistance changed the task sufficiently to make it substantially easier. For another example, there were cases where "Sometimes, I would cue up a praise response, only to have the RT negate it by requiring the subject to repeat the phrase." (Barry Kort, personal communication, November 9, 2001). In these cases, the Reading Tutor and the wizard were giving mixed messages – presumably leading to a less engaging experience than one where the praise and the task structure matched. Such clashes between the Reading Tutor and the human tutor help reveal the limitations of our current implementation of this "Reading Tutor plus" methodology. Further tweaking of the Reading Tutor installation for these tasks – wordlists and limericks – and a tighter integration of the human-generated responses into the dialog management of the Reading Tutor, ought in principle to alleviate such problems.

6 Future Work

We noted that students experienced frustration at times with the Reading Tutor. From the research viewpoint, this frustration represents both threat and opportunity. It's a threat to the study to the extent that it dominates affect and swamps the effects we're trying to study. It's an opportunity to the extent that generating frustration lets us study emotional scaffolding methods for overcoming it – or reveal areas on which to work towards the goal of modifying the system to be less frustrating.

As a reviewer noted, "Although the human wizard is pretty constrained in when and how he can interact with students, he still has significant freedom in what he chooses to say and when he chooses to intervene, so it isn't clear that the authors will be able to automate the types of things the wizard is doing." Further experiments might be directed at finding out what an emotionally supportive automated tutor should say – in the face of limited sensing, or when enabled with richer senses including vision processing.

We found logistics to be a challenge, due to physical separation between where the researchers were based and where we had a partnership with an elementary school. Later experiments might modify the methodology – perhaps separating the wizard from the student not just by a television connection to the next room, but by a Webcam so that the student and the wizard could be anywhere, or at least anywhere with a fast Internet connection.

Future work towards specifying exactly what emotional scaffolding is and how and when it helps could proceed by transcribing and analysing the videotaped interactions between the Reading Tutor, the student, and the human wizard for content and effectiveness. As one reviewer pointed out, "Clearly the biggest area for future work is understanding the human's utterances and when they are used so that they can be automated."

Future work towards improving the Reading Tutor in terms of the affective quality of children's experience might take several directions. If we can understand when kids get confused, we may be able to modify the design to reduce confusion or add suitable prompts. Improving speech recognition accuracy would help a lot, and we're working on it, but it won't happen overnight. If kids get distracted because they have to wait too long for the Reading Tutor to prepare for its next action, or if it takes too long to detect when they reach the end of the sentence, we may need to run on a faster platform, or modify the code. If the limericks in this experiment bore kids enough to be a problem, we may need more interesting materials to use when conducting later experiments – but it may be difficult to give them the nicely controlled experimental properties that the limericks provide.

Future work towards a model of emotions and learning that works for children's reading would need to take into account the intrinsic reward structure of reading. Learning to read may simply not involve lots of insights. The one big aha! is grasping the alphabetic principle – "cracking the code" in the sense that there exists a code. The rest may simply be lots of practice to acquire lots of specifics – particular letter-sound mappings, and meanings of thousands of specific words. That is perhaps a very different set of intrinsic rewards than a set of problem-solving tasks, each task with its own "aha!".

7 Conclusion

We summarize this paper's contributions as follows.

First, we presented a *novel finding on the role of emotions in learning*: Human-provided emotional scaffolding in the context of an intelligent tutoring system for oral reading leads to increased student persistence. This paper presents the first statistically reliable empirical evidence from a controlled study for the effect of human-supplied emotional scaffolding on student persistence in an intelligent tutoring system: Children's persistence was higher with emotional scaffolding than without it. As a reviewer of an earlier version of this paper noted, "affective interaction is an essential feature of a much work in animated pedagogical agents. Results from such work have already demonstrated that interventions by animated agents capable of affective intervention yield positive results in terms of persistence and subjective experience. For example, Lester et al. document the persona effect and its role in promoting learning. This study differs in that it attempts to focus specifically on the effects of emotional scaffolding." Thus we submit for the research community's consideration that intelligent tutoring systems ought to incorporate emotional awareness (sensing and responding) into their behavior.

Second, we presented *examples of emotional scaffolding.* We presented some types of specific interventions used by an expert to provide emotional scaffolding shown to increase persistence.

Finally, we presented a *case study in studying ITSs by adding human expert capabilities to existing tutoring systems*. This experiment tested the effects of adding a human expert capability (in this case emotional scaffolding) to an existing intelligent tutoring system with minimal or no code changes. Perhaps this notion may prove useful not only for exploring emotional scaffolding, but other aspects of intelligent tutoring systems as well.

8 Acknowledgements

This work was supported in part by the National Science Foundation under Grant No. REC-9979894 (Project LISTEN), and under Grant No. REC-0087768 (The Affective Learning Companion). 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.

For Project LISTEN's Web site, see http://www.cs.cmu.edu/~listen. For the Affective Learning Companion's Web site, see http://www.media.mit.edu/affect/AC_research/lc/index.html. Early discussions (c. 1997) on supplementing the Reading Tutor with emotional awareness were had with John Kominek.

References

1. Aist, G. Dissertation 2000. Helping Children Learn Vocabulary during Computer-Assisted Oral Reading. Ph.D. dissertation, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. http://www.cs.cmu.edu/~aist/Aist-PhD-dissertation.html
2. Coles, Gerald. 1999. Literacy, emotions, and the brain. Reading Online, March 1999. http://www.readingonline.org/critical/coles.html Downloaded September 10, 2001.
3. Damasio, A. R. 1994. Descartes' Error: Emotion, Reason, and the Human Brain. New York: Gosset/Putnam Press.
4. Isen, A. M. 2000. Positive affect and decision making. In M. Lewis and J. Haviland, eds. Handbook of Emotions. New York: Guilford. 2nd edition.
5. T. del Soldato and B. du Boulay. 1995. Implementation of motivational tactics in tutoring systems. *Journal of Artificial Intelligence in Education* 6(4): 337-378.
6. Kort, B., Reilly, R., and R. Picard. 2001. An affective model of interplay between emotions and learning: Reengineering educational pedagogy – building a Learning Companion. International Conference on Advanced Learning Technologies (ICALT) 2001. August 6-8, 2001, Madison, Wisconsin.
7. Kucera, H. & Francis, W. N. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence, RI.
8. Lepper, M. R., Woolverton, M., Mumme, D. L., and Gurtner, J. 1993. Motivational Techniques of Expert Human Tutors: Lessons for the Design of Computer-Based Tutors. In S. P. Lajoie and S. J. Derry (Ed.), *Computers as Cognitive Tools*. Hillsdale, NJ: Erlbaum.
9. Jack Mostow, Greg Aist, Juliet Bey, Paul Burkhead, Andrew Cuneo, Susan Rossbach, Brian Tobin, Joe Valeri, and Sara Wilson. NAACL 2001. A hands-on demonstration of Project LISTEN's Reading Tutor and its embedded experiments. Refereed demo presented at Language Technologies 2001: The Second Meeting of the North American Chapter of the Association for Computational Linguistics.
10. Mostow, J. & Aist, G. FF 2001. Evaluating tutors that listen: An overview of Project LISTEN. In (K. Forbus and P. Feltovich, Eds.) Smart machines in education. MIT/AAAI Press, 2001.
11. Mostow, J., & Aist, G. CALICO 1999. Giving help and praise in a Reading Tutor with imperfect listening – Because automated speech recognition means never being able to say you're certain. *CALICO Journal* 16(3): 407-424. Special issue (M. Holland, Ed.), *Tutors that Listen: Speech recognition for Language Learning*.
12. Picard, R. 1997. Affective Computing. Cambridge: MIT Press.