Pause the Video: Quick But Quantitative Expert Evaluation of Tutorial Choices in a Reading Tutor that Listens

Jack Mostow, Cathy Huang, and Brian Tobin
Project LISTEN
Robotics Institute, Carnegie Mellon University, Pittsburgh PA 15213 USA
http://www.cs.cmu.edu/~listen

Abstract. To critique Project LISTEN's automated Reading Tutor, we adapted a panel-of-judges methodology for evaluating expert systems. Three professional elementary educators watched 15 video clips of the Reading Tutor listening to second and third graders read aloud. Each expert chose which of 10 interventions to make in each situation. To keep the Reading Tutor's choice from influencing the expert, we paused each video clip just before the Reading Tutor intervened. After the expert responded, we played back what the Reading Tutor had actually done. The expert then rated its intervention compared to hers.

Although the experts seldom agreed, they rated the Reading Tutor's choices as better than their own in 5% of the cases, equally good in 36%, worse but OK in 41%, and inappropriate in only 19%. The lack of agreement and the surprisingly favorable ratings together suggest that either the Reading Tutor's choices were better than we thought, the experts knew less than we hoped, or the clips showed less than they should.

1. Introduction

Evaluation of automated tutors is both important and difficult. Important because tutors are seldom effective at first, requiring many design iterations to fulfill their potential. Difficult because it takes considerable time, money, and work to determine whether an automated tutor helps students learn anything at all, let alone more than a nontrivial control treatment, let alone more than another version of the same tutor – and if so, which kinds of students, what target skills or concepts, and why!

This paper offers a modest shortcut as a partial solution to this problem – not a panacea, but a tool that may help in some cases. It requires human experts who are willing and able to evaluate a tutorial intervention just by observing it. Such evaluation may be less accurate than laborious pre- and post-testing, but gives a useful sanity check. Controlled comparisons of alternative policies with student achievement as the outcome (eg. [1]) remain the gold standard. However, such evaluations are arduous and require a large investment of resources and time. Advantages of the method proposed here include simplicity, illumination of what makes a particular decision good or bad for a particular student or target, and sensitivity to subtle features that more formal methods might not measure.

We present a case study of using this method to evaluate the system that motivated it: Project LISTEN's Reading Tutor, shown in Figure 1 and described in detail elsewhere [2,6]. The Reading Tutor listens to children read aloud, and helps them learn to read [3, 4]. Its richly interactive, multimodal dialog with the student reflects both opportunities and limitations of technology, but is modeled in part after human tutoring. The Reading Tutor's interactions were originally designed based on observations of expert human tutors and Wizard of Oz experiments [5], followed by years of formative evaluation and design iteration [2]. We wanted to complement our months-long controlled studies of the Reading Tutor's overall effectiveness [3, 4] with a more "quick and dirty" evaluation of how reasonably it was behaving, and where it most needed improvement.

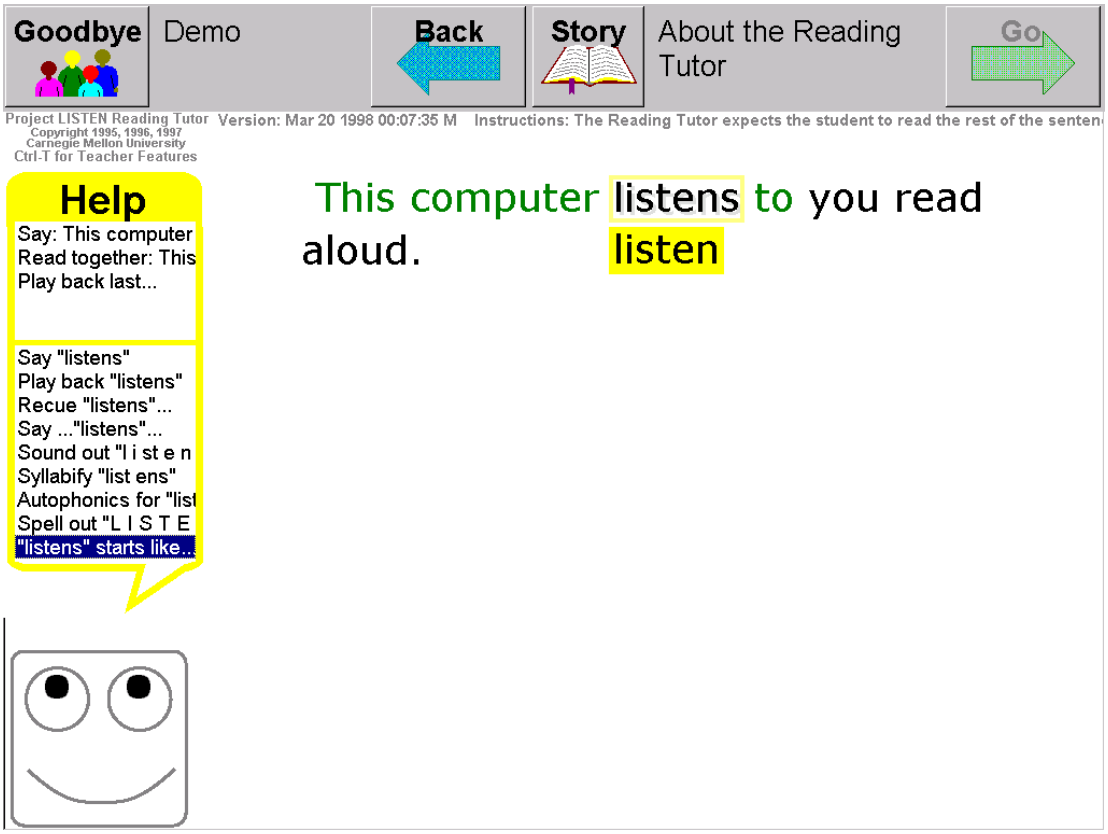

Figure 1: Project LISTEN's Reading Tutor

In particular, we wanted to evaluate the 1999-2000 version of the Reading Tutor's decisions about how to intervene on a given word when the student made a mistake, got stuck, or clicked on the word for help. As Figure 1 illustrates, the Reading Tutor had a repertoire of several kinds of help, such as reading the word aloud, sounding it out, or giving a rhyming hint [6]. The Reading Tutor first computed which types of help were possible on the word in question. For example, a rhyming hint required the word to have a rhyme. The Reading Tutor filtered out hints that were possible to give but considered infelicitous for certain word lengths, word types, or at certain sentence locations. For example, sounding out a word is reported to be ineffective for words longer than four phonemes. Finally, the Reading Tutor simply chose at random from the remaining subset of interventions. This policy provided some variety and served as a plausible baseline behavior in the absence of further knowledge. However, it certainly left room for improvement of the decision method for determining when to give which type of help. Before focusing considerable resources in that direction, we wanted to determine its likely payoff. We therefore decided to compare the Reading Tutor's (random) decisions to those

of professional educators. By asking experts to critique the Reading Tutor's decisions, we sought to not only find out which decisions were bad, but also how to improve them. This paper tells how we proceeded and what we learned.

2. Pause the Video Method

We adapted the "panel of judges" method of expert system evaluation. There are various versions of this method [7, 8, 9, 10], but the key idea is to give multiple human experts the same concrete problems to solve and evaluate the quality of the system's decisions compared to their own and each other's – ideally without them knowing whose decisions are whose. If the experts agree with the system as much as they agree with each other, the system's decisions (and the system) are considered effective.

Our experts were three educational professionals, albeit with varying degrees of expertise and experience in teaching reading. The first teacher was a credentialed elementary school teacher with training and experience in teaching language arts and assessment. The second teacher had experience primarily teaching math and science, with little formal background in reading, but had herself learned English as a second language, and had been paying attention to reading instruction issues for second language learners. The third teacher was a special education teacher with a doctoral degree and experience as a social worker and counselor.

We interviewed each expert separately in a room with a VCR and a television. We explained that we were interested in exploring the appropriateness of the Reading Tutor's different types of help. We gave each participant a twenty-dollar gift certificate to the university store afterwards as compensation for their time.

In our case, the concrete problems presented to the experts consisted of choosing what type of reading help to give the student in a given situation. To depict the situations with sufficient detail, we presented them in the form of digital video clips from actual Reading Tutor sessions in classrooms. We selected video clips of second and third grade children- both boys and girls- reading various passages on the Reading Tutor. Each video clip showed the computer monitor as the child read and a mirror that reflected the child's facial expression. This setup allowed the experts to see what the child saw as well as visual cues from the student such as the student's attentiveness, expression, and gender.

We now describe the successive steps of the Pause the Video method used in this study.

2.1 Introduce the types of help

To orient each expert to the task, we gave her a booklet with these instructions:

We're trying to find out whether Reading Tutor help on specific words is appropriate. First you will be shown the 10 different types of help the Reading Tutor offers on words.

Next you will see a total of 15 pairs of video clips. The first clip starts when a sentence first appears up until the point before the student gets help on a word. You will be asked to pick the best Reading Tutor response from the 10 different types of help. The second clip shows the Reading Tutor giving help on a certain word. You will be asked to rate the Reading Tutor help. Below each answer you have space to explain why. The students will be either 2nd or 3rd graders.

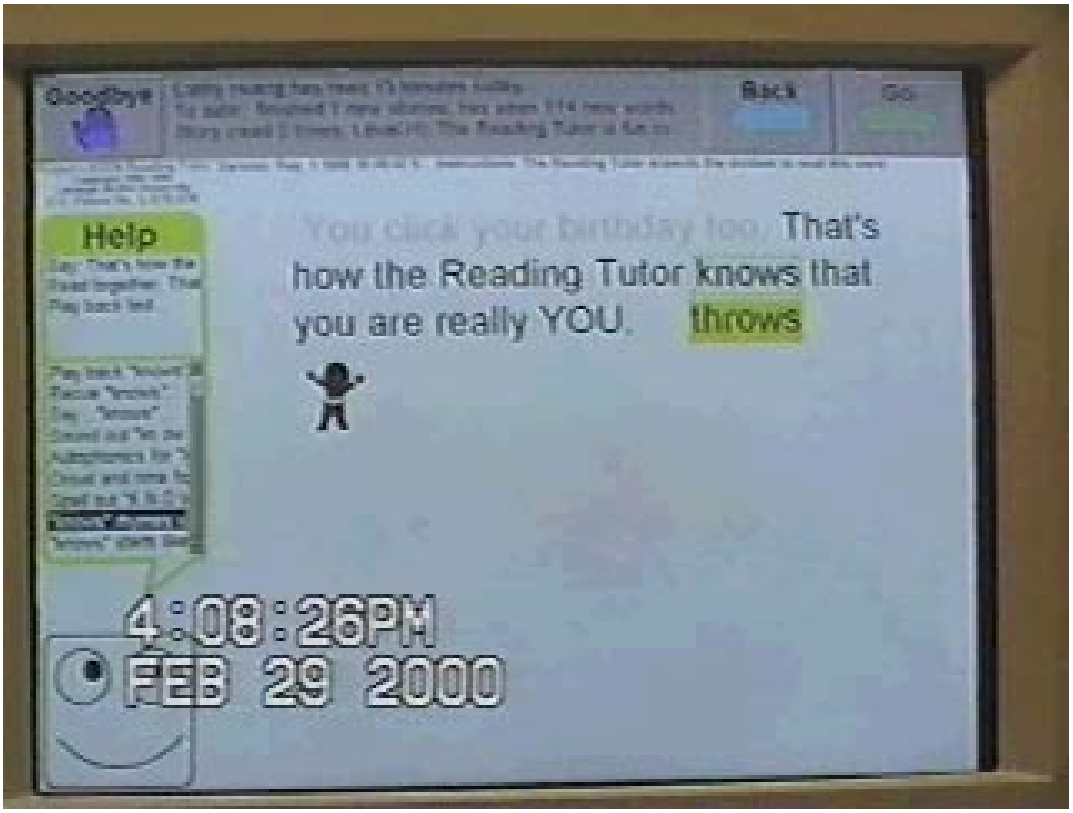

Figure 2: Example intervention used to orient experts

We then played ten video clips, each showing a different type of Reading Tutor help. In the example in Figure 2, the Reading Tutor gives a rhyming hint for the word *knows* by displaying the word *throws* beneath it, saying "rhymes with *throws*."

We provided the following guide for reference throughout the rest of the protocol:

Autophonics
RT highlights "animal" and says "M here makes the sound mmmm" (phonics on specific letter)

OnsetRhyme
RT highlights "zoos" and says "zzz oooz"
(the onset is the first consonant cluster, if any, in a syllable; the rhyme is the vowel and any subsequent consonants)

Playback
RT plays back last recording

Recue
RT says "Many kinds of..." and underlines "animals"

RhymesWith
RT highlights "kinds", displays "finds", and says "Rhymes with "finds"

Say
RT says word

SoundOut
RT highlights "many" and says "M EH N EY" (one phoneme at a time)

SpellOut
RT highlights "zoos" and spells out "Z O O S"

StartsLike
RT highlights "animals" and says "Starts like animal"

Syllabify
RT highlights "animals" and says "a ni mals"

2.2 Pause the video for expert to choose appropriate help

We showed each expert 15 video clips of children reading stories on the Reading Tutor. Each clip showed a child having difficulty with a different word. In the example shown in Figure 3, the reader has encountered the word "Hound."

We then asked the expert which of the ten types of help she considered best in that situation, and why. To avoid biasing the expert's recommendation, we paused the video clip just before the point where the Reading Tutor intervened. To provide what we hoped was adequate context on which to base an informed decision, each video clip started several seconds before the Reading Tutor intervention. We did not provide additional information about the students other than what grade they were in. The expert answered on the form shown below Figure 3.

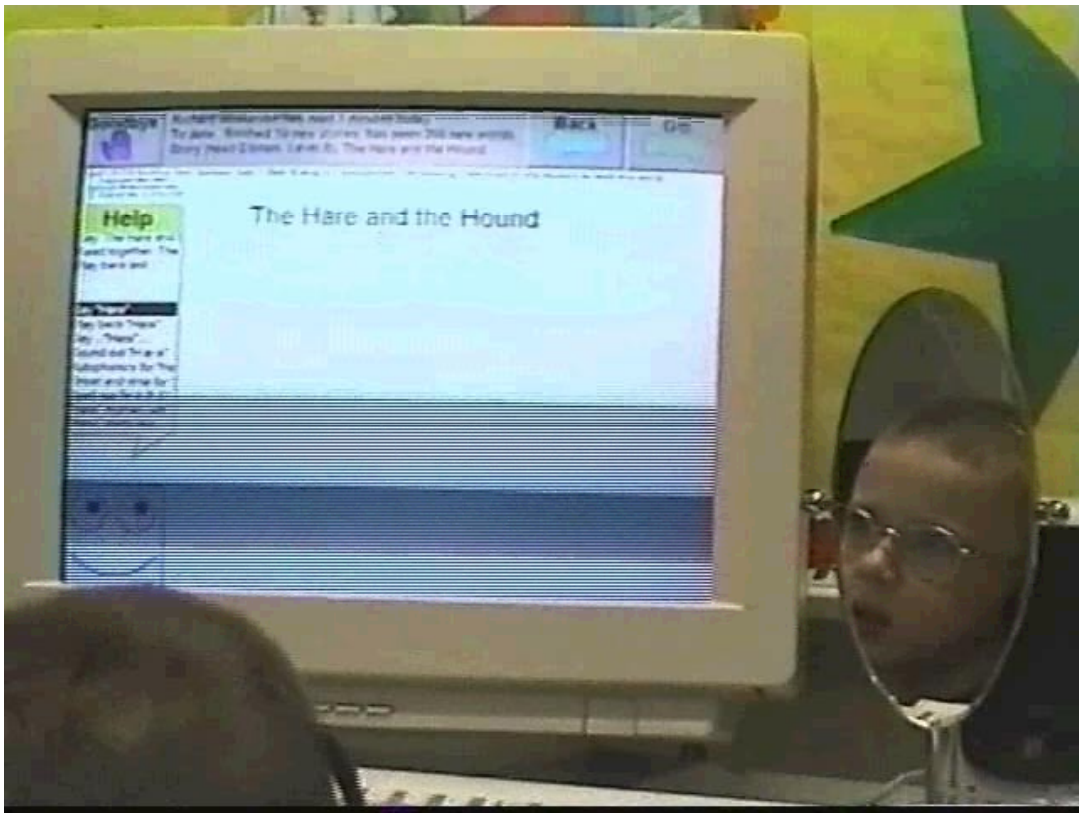

Figure 3: Occasion for tutorial intervention

Child: "the hare and the…"

Clip 6 (2nd grader): The Hare and the **Hound**

Which type of Reading Tutor help is the best?

Autophonics ("S here makes the sound zzzz")

Onset and rhyme ("nnn ohs")

Playback (RT plays back last recording of student's voice)

Recue (RT reads from start of sentence up until word)

Rhymes with (“rhymes with rows”)

Say (RT says word)

Sound out (“nnn oh zzz”)

Spell out (“k n o w s”)

Starts like (“starts like known”)

Syllabify (“tu tor”)

Why?

If different from Reading Tutor help, what would you do?

Step 3: Play back the actual intervention and ask the expert to rate it

After the expert chose a response, we asked her to rate the Reading Tutor’s choice. We played back a follow-on video clip showing how the Reading Tutor had actually responded. The expert rated the type of help chosen by the Reading Tutor on a scale from 1 to 4, and could write down her reason for the rating, using the form shown in Figure 4.

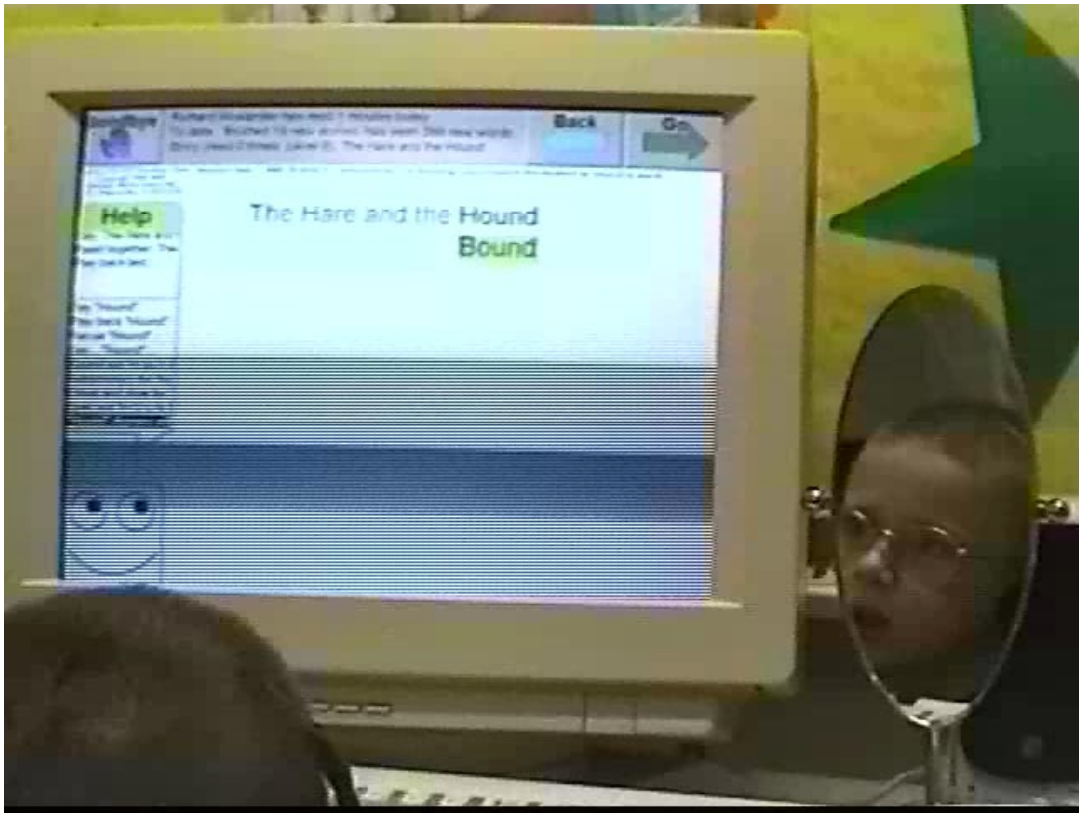

Figure 4: Reading Tutor intervention

Reading Tutor: “rhymes with *Bound*”

Clip 6 (2nd grader): The Hare and the **Hound**

The Reading Tutor gave this type of help: Rhymes with

How does the Reading Tutor help compare to your choice of help?

better

equally good or same

worse but OK

inappropriate

Why?

3. Results

We had asked two questions about each of the 15 examples: which type of Reading Tutor help is the best, and how good is the type Reading Tutor chose? We now analyze the experts' responses, but first we address a more basic issue: did our setup even make it possible to answer the questions?

3.1 Adequacy of the setup

First we wanted to know if the experimental setup enabled the participants to answer the questions at all – by no means a foregone conclusion! As one of the reviewers put it: "If the teacher sees only a single clip for a child, there is scant opportunity to glean any information even about the general level of skill of the child, let alone any information about style, motivation, etc."

Were the experts willing and able to answer the questions based on the amount of information provided in the examples? Yes. All three experts answered all the questions. In rating the Reading Tutor's interventions, the first two understandably wanted to see the students' response to the help to see how they fared, but this information would have biased the ratings to reflect information unavailable when the tutor must decide what to do. Otherwise, none of the written comments complained that the setup provided insufficient information to make an informed answer. This result is consistent with findings that teachers, rather than forming an extensive diagnostic model of the student, tend to follow a "curriculum script," "gathering enough information from student performance cues to correct any student difficulties or misconceptions that might arise" [11, p. 17].

Did the 10 listed choices include the recommended interventions? Yes. For 14 of the 15 examples, all three experts chose one of the listed interventions. The exception was an example that involved the word *goat-herd*. The first expert chose none of the options, writing instead "RT needs to re-read sentence very slowly – it is telling the student s/he is not fluent. This can be intimidating. Re-read slowly, have student repeat each word after me." The second expert wrote "see below" next to the choice *Say*, and "goat-herd is 2 words." It should be noted that the answer form did not explicitly list "none of the above" as a choice, just the question "If different from Reading Tutor help, what would you do?"

Given more information or options, the experts might have answered differently. As the same reviewer went on to suggest, "Varying the amount of lead-in, plus perhaps using the one child for more than one clip, would allow study of what information about the learner the skilled teacher draws on to formulate their comments. (What is the teacher's learner model?)" But the experimental setup succeeded at least to the extent of appearing to work for 14 questions, the answers to which we analyzed.

3.2 Choosing what type of help to provide

How often did the experts choose the same intervention as the Reading Tutor – or each other? Table 1 shows the Reading Tutor's and experts' choices of help for each example.

The experts preferred certain types of help. *Sound Out*, *Say*, *Syllabify*, and *Rhymes With* together accounted for 95% of the experts' responses. We compared each subject's use of these four most frequent types of help, and collapsed the other six less-frequently used

types of help into an *Other* category. During analyses, we kept in mind the fact that this change would inflate the Kappa values slightly if two experts chose different interventions in the Other category for the same example – but as Table 1 shows, that never happened.

How often did the experts agree with each other on what type of help to give? Rarely: Kappa (n = 42) = 0.103, p < 0.156. (Here, 0.103 measures agreement – very slight – and 0.156 is the probability of that agreement occurring by chance.) For 6 of the 14 examples (43%), two of the three agreed. On only one example (the same example shown in Figure 3) did all three choose the same intervention.

How often did the Reading Tutor's choice match the experts'? Rarely; they seldom matched each other's! The Reading Tutor agreed with at least one human subject on five examples (36%), and with all three experts on one example (7%). Agreement between the Reading Tutor and the three experts on when to use *Sound Out*, *Say*, *Syllabify*, and *Rhymes With* was only at chance level, Kappa (n = 42) = 0.029, p < 0.599.

Table 1: Comparison of interventions chosen by Reading Tutor and experts

			Video Clips Shown (caps indicates target word)	Reading Tutor Help	Expert 1 Help	Expert 2 Help	Expert 3 Help
Average Rating of the Reading Tutor 4 = Better than me 3 = Same as mine 2 = Worse but OK 1 = Inappropriate	1.67		A Doe hard pressed by HUNTERS	Sound Out	Other	Syllabify	Say
			...those beds of bright flowers and those cool FOUNTAINS	Sound Out	Rhymes with	Syllabify	Syllabify
			for she had read SEVERAL	Sound Out	Recue	Syllabify	Starts like
	2.00		Now it happened one day that he had an AUDIENCE	Say	Say	Syllabify	Sound Out
			and in order to appear a PERSON	Say	Playback	Syllabify	Syllabify
			and in order to appear a person of some IMPORTANCE	Say	Sound Out	Syllabify	Sound Out
			tale can be a kind of ACCOUNT	Sound Out	Other	Syllabify	Starts like
			All 17 species of penguins live in the Southern HEMISPHERE	Say	Say	Syllabify	Sound Out
	2.33		...predators where their inability to fly is not DETRIMENTAL	Say	Say	Syllabify	Syllabify
	2.67		A Doe hard pressed by hunters SOUGHT	Rhymes with	Onset and Rime	Say	Rhymes with
			Or, we can write it as O O	Say	Spell out	Say	Autophonics
	3.00		A Doe hard PRESSED	Say	Sound Out	Rhymes with	Rhymes with
			The Hare and the HOUND	Rhymes with	Rhymes with	Rhymes with	Rhymes with
			look SOOT	Onset and rime	Sound Out	Sound Out	Rhymes with

3.3 Rating the Reading Tutor

How did the experts rate the Reading Tutor's choices? Surprisingly well, in light of their randomness. Table 1 shows the average rating of the Reading Tutor's choice for each example, ordered from worst to best. Of 42 replies to "How does the Reading Tutor help compare to your choice of help?," 2 (5%) were "better," 15 (36%) were "equally good or same," 17 (41%) were "worse but OK," and only 8 (19%) were "inappropriate." (For the excluded *goat-herd* item, two were "equally good or same," and one was "worse but OK.")

How well did the experts' ratings agree? A bit surprisingly, no different than chance: Kappa (n = 42) = 0.069, p < 0.475. That is, the experts generally rated the Reading Tutor's choices as OK or better, but disagreed otherwise. The low Kappa value may be due in part to treating the ratings as categorical rather than ordinal.

4. Conclusions

This work suggests lessons both for the Reading Tutor and for the Pause the Video method.

4.1 What did we learn about the Reading Tutor?

Compared to conventional summative evaluations of automated tutors [2, 3, 4], the Pause the Video method provided a quick and easy evaluation of the Reading Tutor. The video clips showed the experts children using the Reading Tutor in its regular classroom setting. It let the experts choose what type of help they would provide to students, and rate the actual interventions the Reading Tutor made. It gave us an estimate of how well the Reading Tutor was choosing, and an assessment of how well and where the experts agreed and disagreed on what type of help to provide. It elicited explanations of experts' thought processes in choosing a type of help, a rough profile of where the Reading Tutor seemed to work, and suggestions for what to fix or add to improve the Reading Tutor's efficacy.

Our goal in using the Pause the Video method was to evaluate and refine the Reading Tutor's randomized decisions about how to help a given student on a given word in a given context. Although the experts seldom made the same choice, they rated the Reading Tutor's decisions better than we expected, classifying them as inappropriate only 19% of them time. Evidently the Reading Tutor's decisions were not clearly awful.

However, the experts disagreed with each other not only in choosing a particular type of help to give, but in rating the Reading Tutor's choices. If the expert choices truly represent the best tutorial decisions possible, and there is no clearly shining best choice, then refining the Reading Tutor's rules for deciding when to provide what type of help might not be worthwhile. The current approach – exclude clearly infelicitous interventions, then choose randomly among the rest – might be about the best we could do. Rather than trying to make better choices about what help to give on individual words, we should then invest our resources in other aspects of tutoring – such as which text to read.

4.2 Why did the experts disagree?

The conclusion above hinges on the lack of expert agreement. Why was agreement so low? One possibility is that some of the experts were not really experts, and lacked the knowledge required to make good tutorial choices. A related hypothesis is that their answers were comparably good, but reflected differences in their training or background.

Another hypothesis is that the experimental setup lacked sufficient information to make informed choices. One type of such information might include diagnostic assessment of the student's reading skills based on instruments designed to measure them, or on previous observation. Another type of information is the history of the student's responses to earlier tutorial interventions: what is the batting average for each type of intervention for that particular student?

A more technical possibility involves the measurement of agreement. For example, the Kappa analysis ignores the fact that some choices and ratings are more similar than others. Also, having the experts rate all the choices – not just the Reading Tutor's – although time-consuming, would have yielded 10 times as many individual ratings with which to assess agreement among the experts about the same 15 examples.

4.3 What is Pause the Video good for?

Pause the Video combines the ecological validity of observing live tutoring with the ability to quantify evaluations by multiple experts of reproducible stimuli, namely video clips. Potential applications of this approach vary in their goals.

Our goal was to elicit expertise in order to evaluate and improve a complex, interactive tutor. For this purpose, the teachers were domain experts from whom to acquire knowledge, and Pause the Video gave a way to acquire that knowledge. The teachers' overall rating of the Reading Tutor's interventions gave us a quick assessment of its

decision-making. Their lack of agreement suggested some limitations on how much we could improve those decisions, albeit based on the rather strong assumptions discussed above. Their detailed comments gave us constructive suggestions for improving decisions.

An inverse goal is to train better human tutors. Simulated students have been studied for this purpose, using computer-generated abstractions of student behavior [11]. Video clips of real students could capture important aspects omitted from such abstractions, and allow tutors-in-training to compare their responses with each other and with expert tutors.

Yet another goal is to study how teachers reason. Videotapes of tutoring sessions has been used to stimulate the recall of the tutors themselves in order to inquire into their reasoning by eliciting what they were thinking at the time [11]. Pause the Video differs by using the same video clips to elicit choices from multiple tutors. Consequently, it may give researchers a tool to quantify the impact on tutorial decisions of knowledge (by comparing differently trained tutors) and information (by varying the amount and kind presented).

Acknowledgements

This work was supported in part by the National Science Foundation under Grant Nos. CDA-9616546 and REC-9979894. 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. We also thank the students and educators at Centennial Elementary School, Gregory Aist for his recommendations on the design of this study, the rest of the Project LISTEN team, and the anonymous reviewers for their helpful comments.

References

[1] A. Corbett and H. Trask. Instructional interventions in computer-based tutoring; Differential impact on learning time and accuracy. *Proceedings of the ACM CHI'2000 Conference on Human Factors in Computing Systems*, 2000, pp. 97-10.

[2] J. Mostow and G. Aist. Evaluating tutors that listen. In K. Forbus and P. Feltovich, (eds.) *Smart Machines in Education: The coming revolution in educational technology.* MIT/AAAI Press, 2001. In press.

[3] J. Mostow, G. S. 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.

[4] J. Mostow, G. Aist, P. Burkhead, A. Corbett, A. Cuneo, S. Eitelman, C. Huang, B. Junker, C. Platz, M. B. Sklar, and B. Tobin. A controlled evaluation of computer- versus human-assisted oral reading. *Proceedings of the Tenth Artificial Intelligence in Education (AI-ED) Conference*, San Antonio, Texas, May 2001.

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

[6] 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.

[7] K. P. Jantke, R. Knauf, and T. Abel (1997). The Turing test approach to validation. In Takao Terano (ed.), 15th International Joint Conference on Artificial Intelligence, IJCAI97, Workshop W32, Validation, Verification & Refinement of AI Systems & Subsystems, August 1997, Nagoya, Japan, pages 35-45.

[8] R. Knauf, A.J. Gonzalez & K.P. Jantke. Validating rule-based systems: A complete methodology. *IEEE SMC '99 Conference Proceedings*, 1999 IEEE International Conference on Systems, Man, and Cybernetics, vol 5, 1999. 744-749.

[9] R.M. O'Keefe and D.E. O'Leary. Expert system verification and validation: A survey and tutorial. *Artificial Intelligence Review 7* (1993), 3-42.

[10] J.A. Wise and M.A. Wise. Basic considerations in verification and validation. In: J.A. Wise, V.D. Hopkin, and P. Stager (eds.), *Verification and Validation of Complex Systems: Human Factors Issues*, vol. 110 of NATO ASI Series, Series F: Computer and Systems Sciences, (1993) pp. 87-95.

[11] R. T. Putnam. Structuring and adjusting content for students: A study of live and simulated tutoring of addition. *American Educational Research Journal 27*:1, Spring 1987, 13-48.