Tech., Inst., Cognition and Learning, Vol. 1, pp.61-81
Reprints available directly from the publisher
Photocopying permitted by license only

Automatically Assessing Oral Reading Fluency in a Computer Tutor that Listens

JOSEPH E. BECK*, PENG JIA[1] AND JACK MOSTOW

Project LISTEN, RI-NSH 4215, 5000 Forbes Avenue
Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA
http://www.cs.cmu.edu/~listen

Much of the power of a computer tutor comes from its ability to assess students. In some domains, including oral reading, assessing the proficiency of a student is a challenging task for a computer. Our approach for assessing student reading proficiency is to use data that a computer tutor collects through its interactions with a student to estimate his performance on a human-administered test of oral reading fluency. A model with data collected from the tutor's speech recognizer output correlated, within-grade, at 0.78 on average with student performance on the fluency test. For assessing students, data from the speech recognizer were more useful than student help-seeking behavior. However, adding help-seeking behavior increased the average within-grade correlation to 0.83. These results show that speech recognition is a powerful source of data about student performance, particularly for reading.

Keywords: Assessment, external validity, reading fluency, curriculum-based assessment, validation, Project LISTEN, Reading Tutor, children, oral reading, speech recognition, student modeling.

1 INTRODUCTION AND MOTIVATION

This paper describes a way to assess students using data captured by automated tutors in the course of their normal use, instantiating the "paradigm for *ecologically valid, authentic, unobtrusive, automatic, data-rich, fast, robust,* and *sensitive* evaluation of computer-assisted student performance" proposed in (Mostow & Aist, 1997). This vision is supported by a Research Council report (Pellegrino, Chudowsky, & Glaser, 2001):

*Corresponding author: Joseph.Beck@cs.cmu.edu
[1]Now at Intermedia Advertising Group (IAG), 339 Broadway, Suite 200, New York, NY 10013, USA

> *"One can imagine a future in which the audit function of external assessments would be significantly reduced or even unnecessary because the information needed to assess students at the levels of description appropriate for various external assessment purposes could be derived from the data streams generated by students in and out of their classrooms." (p. 284)*

The ability to continuously and automatically assess students has obvious appeal. The report goes on to compare paper tests with shutting down a store in order to take inventory. Stores are in business to sell goods, not to take inventory. Advances in technology with bar codes, automatic scanners, and computers enable businesses not only to avoid constantly counting how much inventory they have, but also to add new capabilities such as monitoring subtle trends in customer purchases. Similarly, schools are primarily in the business of teaching kids, not taking inventory of how much learning has occurred.

This paper demonstrates, in the context of Project LISTEN's computer tutor for reading, that a similar technological shift for schools is possible, at least within the context of the computer tutor we studied. A challenge in attaining this vision is to map student behaviors to a "knowledge inventory" of what the student knows.

Project LISTEN's Reading Tutor (Mostow & Aist, 2001) is a computer tutor that listens to children read aloud and provides feedback. It is more difficult for a computer to assess student performance in reading than in (for example) mathematics; computers can evaluate student responses to mathematics question, and can sometimes determine where students made mistakes (Anderson, Boyle, Corbett, & Lewis, 1990; Burton, 1982). It is harder for a computer to judge a student's oral reading than it is to evaluate typed input. However, humans are capable of assessing a variety of reading skills by listening to students read words aloud. For example, the Woodcock Reading Mastery Test (Woodcock, 1998) enables human scorers to judge a student's ability at identifying and decoding words, and even to perform an item-level analysis to look for systematic mistakes. The difficulty is in enabling a computer to approach these capabilities. Our approach is to use automated speech recognition technology (Huang et al., 1993) that listens to students read aloud.

Our method is to find properties of the student's reading, that our speech recognizer can detect, that relate to a student's proficiency in reading as measured on human-administered tests. Relating student performance within the tutor to his performance in an unassisted test environment allows for tutor

claims about the student's proficiency that have meaning outside of the context of the tutor. For example, it is more meaningful to a teacher to know that the tutor estimates a particular student's fluency as 35 words per minute than to know that the tutor estimates the student's fluency score as 0.43 on a scale of 0 to 1 defined only with respect to the tutor. Furthermore, the external test serves as a means to verify the accuracy of the tutor's estimate.

We use inter-word latency (Mostow & Aist, 1997) in this paper as an automated measure of student reading ability. Inter-word latency (or simply 'latency') is the time that elapses between reading successive text words, including "false starts, sounding out, repetitions, and other insertions, whether spoken or silent." Mostow & Aist (1997) investigated latency for eight students who used an earlier version of the Reading Tutor over the 1996-97 school year. They found that the mean latency for 36 stop words[2] was lower than the mean latency for non-stop words, and that latency decreased significantly over time from a student's first to last encounter of a word.

However, Mostow & Aist (1997) did not relate latency to established performance measures. Therefore, while proposing latency as a measure of reading proficiency is reasonable, we are not certain whether changes in latency are related to verifiable claims about the student's reading proficiency. Some previous research has studied the relation between time-based measures of student reading and other reading assessment measures. De Soto & De Soto (1983) used data from 134 fourth graders and found a significant negative correlation between the time to read a word and reading comprehension. They used a timed reading of a word list to infer a student's average time to read a word in isolation. In contrast, our latency measure applies to whatever connected text each student encounters in the Reading Tutor. The goal of this paper is to use automated latency measurements to help predict the fluency of students' independent oral reading, and to validate predicted against actual fluency.

2 APPROACH

In this section, we first give the precise definition of inter-word latency and describe the subset of latencies we chose to use in this study. We then describe the dataset we used.

[2]Stop words are a, all, an, and, are, as, at, be, by, for, he, her, him, his, I, if, in, is, it, its, me, not, of, off, on, or, she, so, the, them, then, they, this, to we, you.

2.1 Definition of inter-word latency

The Reading Tutor presents reading material one sentence at a time on the computer screen. While the student reads, the Reading Tutor listens and aligns the speech recognizer output to the actual sentence text. The version of the Reading Tutor in this study used the same algorithm as (Mostow & Aist, 1997) to align each word of text to at most one word in the speech recognizer output. We use the aligned output for our computation of latency.

The speech recognizer could decide the student read the word correctly; that he misread it; or that the student did not attempt to read the word. Given the inaccuracy of speech recognition, the three cases are not the same as "the student read the word correctly; he misread it; he did not attempt to read it." The Reading Tutor is quite good at accepting correct reading (97% accuracy), but detects only about a quarter of the cases where a student misreads or mispronounces a word (Banerjee, Beck, & Mostow, 2003), or about half the reading mistakes serious enough to threaten comprehension (Mostow, Roth, Hauptmann, & Kane, 1994). The inter-word latency for a text word, the i^{th} word in the displayed sentence to be read, is defined as follows:

I If w_i was accepted as correctly read by the recognizer starting at time $t_{i,\,start}$

II And if w_{i-1} was heard (either as being read correctly or incorrectly) at time $t_{i-1,end}$

III Then, the inter-word w_i is $t_{i,start} - t_{i-1,end}$

IV Otherwise, the inter-word latency is not defined.

Table 1 provides an example. If the actual sentence text was: "It was the worst quake ever" and the recognizer heard "*it…the were quake ever*," then the algorithm marked word "was" as skipped and the word "worst" as misread, and accepted all other text words as correct.

Thus, inter-word latency is defined only for words that are believed to be read correctly, and only if the previous word in the sentence was not omitted by the student. Therefore, the first word of any sentence will never have a latency (e.g., the word "it" in Table 1). There is no latency for the word "worst" because it was not read correctly. The word "was" does not have a latency because (according to the speech recognizer) it was omitted by the student. Therefore, the word "the" has no inter-word latency, since the word preceding it was omitted (the second condition in the latency definition). The other two

TABLE 1
Latency computation

Sentence text	Speech recognizer output for each word	Start time (ms)	End time (ms)	Latency (ms)
It	IT	0	360	-
was		480	610	-
the	THE	650	860	-
worst	WERE	1040	1140	-
quake	QUAKE	1350	1510	210
ever	EVER	1570	1640	60

words in the sentence have latency measures as shown in the last column in Table 1. For example, for the word "ever," the student started to read the word at 1510ms, and finished reading the word "was" at 1570ms. Therefore, the latency is 1570-1510= 60ms. All times are multiples of 10ms because the recognizer discretizes time into 10ms frames. Each hypothesized word covers a sequence of frames. Thus two words recognized in succession, with no intervening pause, would have inter-word latency measured as 10ms.

Furthermore, latency is defined as the time from when the student finishes saying word *i-1* until he begins to say word *i* correctly. For example, suppose the student said "We are leaving to…tuh….tomorrow." The latency for the word "tomorrow" would include the time spent saying "to…tuh…" Thus, the latency includes the time students spend making false starts towards pronouncing the word. The goal is to estimate how long it took the student to identify the word, whether silently or noisily; this quantity should be a sensitive indicator of automaticity in identifying specific words, and of overall oral reading fluency.

2.2 Which latencies to consider?

We considered only the first attempt a student made at reading a sentence, even if he did not read the entire sentence. We approximated "attempt" as "utterance," operationalized by the Reading Tutor's segmentation of its input signal into utterances delimited by long silences. On subsequent attempts to reread the sentence, the student could just repeat from short-term memory the words that he or the Reading Tutor had just read. Since the student would not have to decode these words, their latencies would be artificially shortened.

To control for word difficulty across students, we only considered words the student encountered at least twice while using the Reading Tutor. For each student, we computed the latency for the first time he encountered a word, and defined this quantity as *initial latency*. We excluded data gathered during a 14-day period after each student started using the Reading Tutor since these data might be conflated with students learning how to use the tutor. We defined each word's *final latency* as in (Mostow & Aist, 1997): we computed the latency only for the student's *first* attempt at reading a word on the *last* day that he encountered that word. Scoring subsequent attempts to read the word would suffer from the recency effect described above. We used only those words that had both an initial and a final latency. This pairing ensured that initial and final latencies were estimated from the same distribution of words, and allowed us to compare the means.

Since we expected latencies to differ based on the difficulty of the word, we classified words as either easy or hard. We operationalized easy words as words on the Dolch list (Dolch, 1936). The Dolch list has 220 very frequent words used in children's books, including all 36 stop words used in (Mostow & Aist, 1997), and is often used in reading proficiency studies (Kersey & Fadjo, 1971). Since Dolch words "glue" a sentence's content together, a student must recognize them quickly so as not to impede the comprehension of the sentence (May, 1998). "Hard" words were operationalized as words not on the Dolch list.

We also excluded "words" the student encountered that did not require decoding. For example, the Reading Tutor had spelling activities where the student had to spell a word aloud (e.g., "CAT"). In this case, the Reading Tutor counted letter names, e.g., "C," "A," "T," as encountered "words," but we did not consider them when computing latency. We also removed words expressed in numerical format, such as "7" and "1999."

2.3 Description of study population and available data

The Reading Tutor logged student reading activities in detail, including the speech recognition output (what it believed the student said) for each word. These logs had a rich description of the student's interaction with the Reading Tutor. We parsed and loaded the log files into a database (Mostow, Aist, Beck et al., 2002). In the 2000-2001 school year, we deployed the Reading Tutor in two Pittsburgh area schools. The schools had both been designated by the United States Department of Education as Blue Ribbon National Schools of Excellence, located in identical buildings two miles apart in the same affluent, suburban school district.

Over the course of the year, 88 students in grades one through four (i.e., 6- through 9-year olds) used the Reading Tutor. Students began using the tutor at the end of October and finished using the tutor in early June, nominally every day for 20 minutes. Analyzing usage data revealed that students interacted with the tutor for 18 hours on average over the course of the year, and used the tutor on 70% of the possible days.

Trained examiners administered fluency tests four times during the school year. Pretests were given October and posttests were given in May. Two interim fluency tests were given, in January and in March. To keep students from memorizing the content of the fluency passages, the pretests and posttests used form A of the fluency test, and the interim tests used forms B and C. Tests were individually administered and scored by hand. Each test consisted of three passages that were selected by reading researcher Dr. Rollanda O'Connor. The student's score on a passage was the number of words read correctly in one minute. The student's score for a test was the median of the three passage scores.

Students read passages at their grade level, so, for example, students in grade three read more challenging material than students in grade one. These tests were administered as part of a study to measure the effectiveness of the Reading Tutor compared to independent reading practice (Mostow, Aist, Bey et al., 2002), and were not administered with the goal of creating this assessment approach. It is important to note that these tests were administered outside the context of the Reading Tutor.

There were 37 first graders in the study (15 girls and 22 boys), 18 second graders (9 girls and 9 boys), 17 third graders (7 girls and 10 boys), and 16 fourth graders (6 girls and 10 boys). Speech recognition data for one third grader were lost during the course of the study, for a total sample of 87 students.

3 PSYCHOMETRIC PROPERTIES OF LATENCY

We now discuss relating our latency measure to fluency tests. We estimated the psychometric properties of latency with a subset of 58 students (out of 87) whose data were available at the time. The analyses in this Section were performed with the goal of directing our future research on assessing students, and were not meant to be summative. Due to changes in how we record student interactions with the Reading Tutor, it would be costly to recalculate the results for the full dataset. Therefore, the analyses in this section used a subset of 58 students, while the analyses in the next two Sections used the complete set of 87 students.

TABLE 2
Descriptive statistics for average initial and final latencies.

Words considered	Average initial latencies (ms)				Average final latencies (ms)			
	Mean	Median	Min	Max	Mean	Median	Min	Max
Just Dolch words	494	427	104	1291	400	346	101	1000
Just non-Dolch	588	507	138	1541	473	402	164	1323
All Words	545	485	136	1240	439	395	164	1128

3.1 Reliability and statistical properties of initial and final latencies

Latency for individual words is very noisy due to inaccuracies in speech recognition, and thus cannot be used to make meaningful predictions. Considering the average latencies of all words (or a large sample of words) that a student reads smoothes out the noise and results in a more useful measure. Therefore, we computed the average of all initial latencies and the average of all final latencies for each student, and used these aggregated results. Each student averaged 524 initial/final latency pairs (minimum=114, maximum=1737, median=411 pairs). Among these latency pairs, about 26% (minimum=10%, maximum=39%, median=27%) are for words in the Dolch list.

Table 2 summarizes how students' average latencies varied. A two-tailed paired T-test on each student's mean initial latency and mean final latency shows that final latencies were significantly shorter than initial latencies ($p < 0.0001$). In addition, for both initial and final latency measures, average latencies of each student for non-Dolch words were longer than average latencies for Dolch words ($p < 0.002$). These findings agreed with those reported by Mostow & Aist (1997) in that differences in word difficulty can be detected with latency.

We also studied the reliability of latency by using test-retest methodology. The students' average initial latencies correlated at 0.82 with their average final latencies for non-Dolch words. Thus, latency scores are fairly stable over time for the purposes of ranking students (even though latencies do, in fact, decrease). Computing reliability using the split-halves method (Crocker & Algina, 1986) gives a correlation of 0.79 for average initial latencies and 0.64 for average final latencies. Using the Spearman-

Brown prophecy formula correction (Crocker & Algina, 1986) gives a reliability of 0.88 and 0.78 for average initial latencies and average final latencies, respectively. Therefore, latency is a fairly reliable measure.

3.2 Construct validity

Fluency tests and latency both measure how well students read, but in different ways. Both latency and fluency are time-related measures. The human-administered fluency tests credit only words that the student reads and pronounces correctly. The latency measure is limited by the speech recognizer's accuracy, and considers only words that the Reading Tutor accepted as read correctly. In our fluency tests, the passages read by students were at their grade level. In constrast, the latency measures do not have any guarantees that the students were reading grade-appropriate material, especially since students chose half of the stories they read. However, the Reading Tutor did attempt (Aist & Mostow, 2000) to give students passages that were at their level of reading ability. The Reading Tutor's assistance to students is also a threat to the validity of latency as a measure of fluency. We ameliorate this problem by considering only a student's first attempt at reading a sentence, before he is likely to have asked for help.

Both fluency and latency basically measure the same underlying construct: "how quickly the student reads." Fluency, measured as words read correctly per minute, can be thought of as (the multiplicative inverse of) how long students take to figure out how to pronounce a word, plus the actual time to say the word. For example, if a student requires, on average, 300ms to figure out how to pronounce a word and 200ms to say it, that student takes 500ms to read a word and can read 2 words per second, or 120 words per minute. (This analysis is appropriate for early readers, who read word by word; in more fluent readers, reading words aloud overlaps in time with identifying subsequent words – but can still be interrupted when the reader must decode a difficult word.)

Latency acts as a microscope to allow us to zoom in on the time the student takes to figure out how to pronounce a word, but does not include the time the student requires to actually say the word. Of the two aspects that a fluency test combines – namely, pronunciation and production time – we are more interested in how long a student takes to figure out the pronunciation, which reflects automaticity of word identification, than in how fast he or she speaks it, which may reflect physiological or regional factors less relevant to reading proficiency. Therefore, latency allows us not only to investigate fluency at a finer grain size, but also to focus on the more interesting component of it.

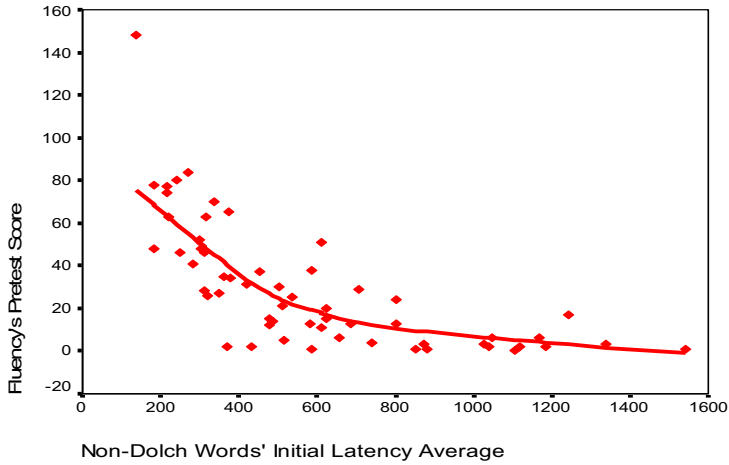


FIGURE 1
Non-Dolch words' average initial latency vs. fluency pretest for each student.

3.3 Statistical validation

To validate latency as a predictor of fluency, we relate a student's pretest fluency score to his mean initial latency, and posttest fluency score to mean final latency. Does (computer-measured) latency correlate with (human-measured) fluency? Figure 1 shows how students' fluency pretest scores relate to their average initial latencies on non-Dolch words. The line is a Lowess curve (constructed by SPSS) generated by fitting 75% of the data. The graph shows a non-linear relationship between the two measures, but it indicates that we might get a linear relation between fluency and latency by taking the inverse of average latency (i.e., 1/latency) for non-Dolch words for each student.

Inverse mean initial latency on non-Dolch words correlated with pretest fluency at 0.86. However, inverse mean final latency for non-Dolch words correlated with posttest fluency at only 0.60. We are unsure why pretest scores were much better estimated than posttest scores.

4 USING SPEECH RECOGNITION TO PREDICT FLUENCY

There were several problems with modeling fluency using initial and final latencies. First, the notion of initial and final does not generalize well to interim times. For example, if we wanted to predict student fluency in February we would be unable to do so. Second, the initial latency for a word

could occur towards the end of the year (and, similarly, a final latency could occur at the beginning of the year), yet such an initial latency would be associated with the pretest. Third, there is a richer source of information available beyond how long the student hesitated before a word. For example, consider a student who has a small sight vocabulary but who cannot decode and does attempt to read complex words. This student might have a lower mean latency than a second student who attempted to read more difficult words, but had to struggle to read some of them. However, the second student would likely have a higher fluency score on a paper test than the first student. Since students decided which words they attempted to read, we are not guaranteed to have a similar sample of latencies across students, even for students reading the same text. Similar considerations suggest that accounting for what percentage of the text and what types of words the student reads may be helpful in assessing students.

To better estimate student fluency, we expand the number of features beyond using just the mean latency. We also use windowing to consider only data that were chronologically close to the test date.

4.1 Features extracted

We construct several features based on latencies and other outputs of the speech recognizer. For each student, we compute the following features:

1. Percentage of words with a defined latency
2. Percent of words with a defined latency measure between 10ms and 5000ms (excludes very fluent reading and possible occurrences of the tutor being confused about where in the sentence the student was reading)
3. Percent of words with latency of 10ms (10ms is the finest grain at which we measure time)
4. Percent of words accepted by the Reading Tutor as correctly read
5. Median of all latencies
6. Mean of all latencies
7. Median of all latencies between 10ms and 5000ms
8. Mean of all latencies between 10ms and 5000ms
9. Percent of words read that were Dolch words

For the first eight items above, we compute the value for all words, for only Dolch words, and for only non-Dolch words (i.e., 3 features per item). Since Figure 1 shows a non-linear relationship between latency and fluency, we also compute the inverse (i.e., 1/x) for the first eight variables.

These additional variables broaden the types of reading phenomena that we can model. For example, word *i* of a sentence had a latency only if word *i*-1 was attempted. The first feature directly measures how much connected reading the student did. If a student's initial attempts to read sentences frequently omitted words, then he should have fewer words with defined latency.

4.2 Modeling approach: windowing

Once we have the features describing the student's performance, we must determine how to use them to predict the student's paper test scores. In order to make assessment dynamic over time (not just a single overall assessment for the entire school year), we consider data about student performance in the Reading Tutor only from within a window of time near when a paper test was administered. If the goal is to use Reading Tutor data to predict how a student will do on an external test, ideally we would examine data from just before the student took a particular test, but such data are not always available. Since the goal of the paper pretest was to measure the reading proficiency of students before they started using the Reading Tutor, the pretests were administered before students started using the tutor. Therefore, we use a window of time *after* students took the pretest. If we restrict ourselves to windows that occurred before the paper tests, we would have no Reading Tutor data from which to predict performance on the first fluency test. Therefore, for the pretest we use a window of time *after* students took the pretest; for the other tests we use a window from *before* students took the test. This non-uniform windowing scheme makes it more difficult to construct an accurate model, and results in models of somewhat lower statistical accuracy, but the model should generalize better to various points in time throughout the school year than a model built without the pretest scores.

Selecting the right window size is difficult. If we are estimating student fluency for March 25, we should give more weight to data from late March than from December. Ideally, a student's performance on the Reading Tutor on March 25 would do the best job at predicting his fluency on March 25. Unfortunately, smaller window sizes result in a less stable estimate of the features; 1000 observations provide a better estimate of the mean than 100. Furthermore, student performance was rather variable, especially when measured with a speech recognition system. For example, perhaps on March 25 the room was noisier than

usual and the Reading Tutor had a hard time hearing the student.

A larger window size is a solution to this problem of variable performance. However, as the window size becomes larger, older student data necessarily are examined. Since students in fact get better at reading over the course of the year, the older data cause a downward bias in estimated reading proficiency. Thus there is a tension between having a small window size to provide a more recent description of student performance, and having a larger window size to reduce the day-to-day variation in student behavior. This bias/variance tradeoff is endemic to many computational modeling problems. We took an atheoretic approach to window size and tried a variety of sizes ranging from one week to three months. There are a variety of techniques for dealing with temporal data. We selected windowing as a first technique to try since it is straightforward to implement, and more complex schemes such as discounting data based on its age have shown mixed results (Webb & Kuzmycz, 1998).

On average, students generated over 100 latencies/day. Students read more than 100 words each day, but we count latencies only from the first attempt at reading a sentence. Furthermore, latency is undefined for the first word of a sentence, and word *i* has a defined latency only if word *i-1* was heard by the speech recognizer.

Once we have a specified window size, for each student we collapse all of his interaction data for the specified window into the set of features described above. Since we also know the student's paper test score, we can relate the set of features to his test score. This approach drops the notion of initial and final latencies. We use all of the latencies within the window (from the student's first attempt at reading a sentence) to compute the features listed above.

Table 3 shows an example of this process. The left two columns indicate which student the data are from and for which test period. We do not use student identity or test date as features, since the goal is to generalize across students and test administration times. We provide the middle columns as features to our model. (Note: This table is abbreviated for space.) The goal of the model is to predict the last column: the student's paper test score.

By ignoring the testing date, we assume that the relation of student performance on the Reading Tutor to performance on paper tests remains constant across the various testing times. The difference in ability to predict pretest and posttest scores (in Section 3.3) suggests that the relation may change over the course of the year. However, for our model we assume constancy. Although this assumption of constancy hurts predictive accuracy somewhat, it enables us to make predictions for months such as February, when no paper tests were administered. We simply determine the date for

TABLE 3
(Abbreviated) Example of features provided to modeler.

Features computed from student-tutor interaction data						
Student ID	**Test date**	**% of words with latency**	**Median latency (ms)**	**Mean latency on Dolch words (ms)**	**...**	**Fluency test score (words per minute)**
30	Oct.	39	310	215	…	45
30	Mar.	52	280	183	…	60
35	Oct.	21	350	221	…	30
35	Mar	56	300	242	…	50
451	Oct	71	290	203	…	40
451	Mar.	80	210	87	…	90

which we want an estimate of student performance, gather data from the preceding window, and feed those data into our model.

Using this assumption of the constancy of the relationship between student performance and paper test scores across the various testing times, we aggregate the data for all 87 students on four fluency tests together, yielding 348 instances with which to train a predictive model of fluency.

4.3 Results for predicting fluency from speech recognizer output

Given the contents of Table 3, we can build a model that takes the student-tutor interaction features and predicts the fluency test score. We use linear regression to build this model. Specifically, we use SPSS's forward regression procedure with P(entry)=0.05 and P(removal)=0.1, and we replaced missing values with the mean. One decision is whether to build a separate model for each grade or a single model for all students. Building a separate model for each grade accounts for the fact that students in different grades took different fluency tests. However, such a model may generalize less well when making predictions outside of the two schools in our study. For example, students in schools in our study were above average in reading proficiency for their grade level.

A regression model for second graders built from this population of students may not generalize to second graders in other populations. Per-grade models are more accurate, but we feel the difficulty in generalizing the results outweighs the gains in statistical accuracy. Therefore, we constructed a single linear regression model for the entire population of students, not one model for each grade. All correlation coefficients we report are for a leave-one-out analysis.

TABLE 4
Correlations for using speech recognition data to predict fluency

	Window size				
	1 week	**2 weeks**	**1 month**	**2 months**	**3 months**
Overall (N=87)	0.63	0.69	0.75	0.79	0.77
Grade 1 (N=37)	0.63	0.68	0.75	0.76	0.78
Grade 2 (N=18)	0.39	0.38	0.52	0.76	0.71
Grade 3 (N=16)	0.73	0.76	0.78	0.79	0.81
Grade 4 (N=16)	0.51	0.61	0.77	0.79	0.74
Mean within-grade correlation	0.57	0.61	0.71	0.78	0.76

That is, for a particular data point, SPSS constructs a regression model using all of the other data and tests the model's fit for that point; this process is repeated for all of the points in the data. Therefore, the correlation coefficients are not overstated and are not the result of overfitting (Mitchell, 1997). All correlations in this section and the next section are significant at $p < 0.01$.

We report within-grade correlations to control for the homogeneity of the population. A heterogeneous population can be well "modeled" by a spurious variable. For example, knowing the student's shoe-size would result in an accurate model of fluency when applied to first through eighth graders (children with small feet are probably younger, and younger children tend to have lower fluencies). However, such a model would perform poorly when applied to only first graders. To ensure we weren't measuring the equivalent of shoe size, we investigated how well our model predicted for each grade.

Table 4 shows the results of this evaluation, both overall and disaggregated by the student's grade. The last row of the table is the arithmetic mean of the within-grade correlations. For example, the mean correlation for a one week window is $(0.63 + 0.39 + 0.73 + 0.51) / 4 = 0.57$. In general, the overall and mean within-grade correlations are very similar.

The within-grade results of the model are fairly strong. For a two-month window, the model accounts for at least 58% of the variance in each grade. Even for a one-week window, it accounts for 16% to 53% of the variation. The regression model generally performs better with a longer window size, although for a 3- month window, results are somewhat poorer than with a 2-month window. For shorter windows, there are considerable differences among grades in the

within-grade accuracy of the model. We are unsure why the model fit for second graders is so poor. However, once the window size reaches 2 months, the model performs nearly identically across grades.

5 ACCOUNTING FOR STUDENT STRATEGY: USING HELP REQUESTS

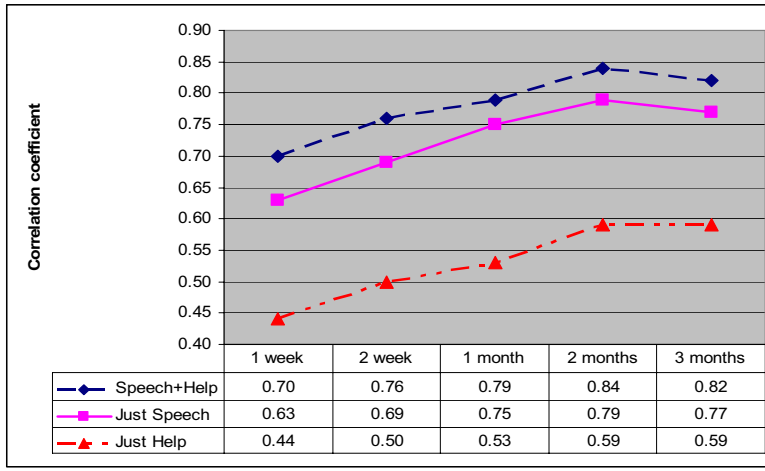
One concern with using speech recognizer information to assess students is that the approach is vulnerable to variations in how students interact with the Reading Tutor. In particular, we know that students' help request rates – the frequency with which they click on words for help – varies from 0.5% to 50% of words seen (Beck, Jia, Sison, & Mostow, 2003). For example, a student who attempts to read each sentence without the tutor's help will not appear as fluent a reader as a student who, before reading the sentence, first asks for help on unfamiliar words in the sentence. When a student requests help it provides us with information about his proficiency, and the help content makes the task easier for the student. For both of these reasons we add to our model information about how often the student requests help.

5.1 Extracting information about student help behavior

Although there has been work on educational data mining in the Reading Tutor (e.g., (Mostow & Aist, 2001)), its logs were not terribly conducive to the process. For example, we know that a student clicked for help on a word, but there isn't a feasible way to compute what type of help the Reading Tutor provided (sounding out a word, saying the word, providing a rhyme, etc.). It was not unusual for students to interrupt hints, clicking repeatedly on a word until the Reading Tutor said it. Therefore we consider only whether the student clicked on the word at all, and do not count how many times he clicked. On average, students clicked for help on roughly 20 words per day. So help requests are a sparser source of data than latencies.

Another problem area was that before a student began to read a sentence containing an unfamiliar word, the Reading Tutor frequently provided preemptive help on the word. Whether the tutor provided such help is known, but the log format made it infeasible to automate the identification of the word on which the tutor gave help. Both issues regarding the recording of help have been addressed in the current version of the Reading Tutor, which logs directly to a database (Mostow, Beck, Chalasani, Cuneo, & Jia, 2002).

We constructed two features to describe student help requests: the



	1 week	2 week	1 month	2 months	3 months
Speech+Help	0.70	0.76	0.79	0.84	0.82
Just Speech	0.63	0.69	0.75	0.79	0.77
Just Help	0.44	0.50	0.53	0.59	0.59

FIGURE 2
Overall correlation between predicted and actual fluency scores, disaggregated by window size and features used.

percentage of words on which the student clicked for help, and the percentage of sentences on which the student requested help. Sentence help consisted of the tutor reading the entire sentence aloud to the student. We followed the same methodology of computing the inverse of the help features as we did for the speech recognition features. We also computed the word help percentage for all words, for Dolch words, and for non-Dolch words.

5.2 Results from adding help requests

We follow the same procedure as in the previous section: we use a forward regression model with leave-one-out cross validation. In addition to the previous model that uses only speech recognition data, we built a regression model that uses features both from speech recognition and from help requests. This model allows us to determine how much additional information is contained in the features describing student help requests. For comparison purposes, we also constructed a regression model that only uses help request data for features. This model serves as a control that allows us to test whether it is worthwhile to use speech recognition to assess students or whether we can accurately assess students just by observing their help-seeking behavior. Figure 2 shows the performance of each of these models across various window sizes. The middle row of the table at the bottom of Figure 2 corresponds to the result for all students in Table 4 .

The combined model of speech recognition data plus help requests substantially outperforms a model that uses only speech data. This improvement ranges from explaining 6% more variance (for a one-month window) to 10% (for a two-week window). The average within-grade correlation is 0.83 for a two-month window using speech data and help requests vs. 0.78 for using only speech data. Therefore, student help requests do contain useful information for assessment. The performance of the model that only uses help requests is not nearly as good as the model that uses speech recognition data. In all cases the model using speech recognition information accounts for at least 20% more variance. In fact, even three months of help request data (approximately 800 help requests on average) does not perform as well as one week of speech recognition data (approximately 300 latencies on average).

6 FUTURE WORK

In the future, the following considerations might help us in better estimating fluency:

1. Finding and correcting problematic latencies. Lengthy latencies might occur in two kinds of situations:

   I Problematic interactions between the student and the Reading Tutor, such as not agreeing on what part of the sentence to read next.

   II When the student is struggling to read the word.

We would like to exclude high latencies of type I. without removing those of type II. Taking these latencies out will help us build better models since we will have cleaner data.

2. Avoiding the use of windows of time. Windows suffer from the bias/variance problem described above, and there is no good *a priori* method to select the best window size. One possibility is to use knowledge tracing (Corbett & Anderson, 1995), which incrementally adjusts a model of student knowledge as new data become available. We have developed a prototype version of using knowledge tracing with the Reading Tutor's speech output (Sison & Beck, 2004), but need to further validate the approach.

Modeling a student's overall fluency is a coarse way to measure reading proficiency. Enhancing our ability to better assess finer grained skills is a logical next step. One approach we have been pursuing is to determine the student's knowledge of subword units, such as the letters "ch" making the sound /k/ as in the word "chaos." We have experimented with using the speech recognizer's judgment to update estimates of the student's knowledge of letter to sound mappings. This work is still in the experimental stage, but success in this area would greatly enhance the diagnostic capabilities of the Reading Tutor.

7 CONCLUSIONS

This paper makes a step towards attaining visions (Mostow & Aist, 1997; Pellegrino *et al.*, 2001) of assessing students based on data streams of their educational activities. We have constructed models that correlate at over 0.7, within grade, with established fluency tests. The within-grade result is particularly good given the difficulty in attaining a strong correlation as the population becomes more homogenous. Neither of the sources of information in this paper, speech recognition output and student help requests, are intrusive or disruptive of the educational process. In fact, this information was being recorded before the research presented here even began. Thus, the students' normal educational activities can occur uninterrupted.

For estimating students' fluency, speech recognizer output contains powerful information. Given that we have only begun to tap the richness in the student's spoken input, and have not used other features (such as pitch), this initial result is encouraging. Eventually, we would like to have estimates of student fluency, using a shorter window size, that are interchangeable with the scores of actual fluency tests.

Using only data available to conventional (i.e., non-listening) tutors, such as student help requests, does not result in nearly as accurate a model as speech recognition data. However, student help requests do have predictive power beyond the information the speech recognizer provides.

The single most useful variable is the (inverse) percentage of words that have a defined latency. This feature is somewhat different from, and outperforms, the percentage of text words that the speech recognizer accepts as correct. Since latency is defined only for two successive words that the student attempts to read, it is not defined for the first word of the sentence or for

isolated words read correctly. Thus, the student's ability to string multiple words in a row together seems to have some predictive power above and beyond just saying those words correctly in isolation.

The approach of enabling a computer tutor to assess a student by relating fine-grained features to existing, external measures is a promising one. In addition to fluency, we have also validated the use of speech recognition data to predict the Word Identification subtest of the Woodcock Reading Mastery Test (Beck, Jia, & Mostow, 2003). Our approach of automatically assessing students by bootstrapping from the extensive effort spent psychometrically validating instruments such as the WRMT both makes it feasible to accurately estimate student reading proficiency and allows the tutor's claims about the student to have more meaning outside the context of the tutor.

Acknowledgements

This work was supported by the National Science Foundation under Grant Numbers 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. We thank other members of Project LISTEN who contributed to this work, and the students and educators at the schools where Reading Tutors recorded data.

REFERENCES

(please see Project LISTEN publications at www.cs.cmu.edu/~listen)

Aist, G., & J. Mostow, (2000, June). *Improving story choice in a reading tutor that listens.* In Proceedings of the Fifth International Conference on Intelligent Tutoring Systems (ITS'2000). 2000. p. 645 Montreal, Canada.

Anderson, J., Boyle, C. F., Corbett, A., & Lewis, M. (1990). Cognitive modeling and intelligent tutoring. *Artificial Intelligence, 42*, 7-49.

Banerjee, S., J Beck, & J. Mostow, (2003, September 1-4). *Evaluating the Effect of Predicting Oral Reading Miscues.* In Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003). 2003. p. 3165-3168 Geneva, Switzerland.

Beck, J. E., P. Jia, & J. Mostow. *Assessing Student Proficiency in a Reading Tutor that Listens.* In Proceedings of Ninth International Conference on User Modeling. 2003. p.323-327 Johnstown, PA.

Beck, J. E., P. Jia, J. Sison, & J.Mostow. *Predicting student help-request behavior in an intelligent tutor for reading.* In proceedings of Ninth International Conference on User Modeling. 2003. p. 303-312 Johnstown, PA.

Burton, R. (1982). Diagnosing Bugs in a Simple Procedural Skill. In D. Sleeman & J. Brown (Eds.), *Intelligent Tutoring Systems* (pp. 157-182): Academic Press.

Corbett, A., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*, 253-278.

Crocker, L., & Algina, J. (1986). *Introduction to Classical & Modern Test Theory.*: Harcourt Brace Jovanovich College Publishers.

De Soto, J. L., & De Soto, C. B. (1983). Relationship of Reading Achievement to Verbal Processing Abilities. *Journal of Educational Psychology, 75*(1), 116-127.

Dolch, E. (1936). A basic sight vocabulary. *Elementary School Journal, 36*, 456-460.

Kersey, H., & Fadjo, R. (1971). Project Report III: A Comparison of Seminole Reading Vocabulary and the Dolch Word Lists. *Journal of American Indian Education, 11*(1).

May, F. B. (1998). *Reading as Communication: To Help Children Write and Read* (5 ed.). Upper SaddleRiver, NJ: Prentice Hall.

Mitchell, T. (1997). *Machine Learning*: McGraw-Hill.

Mostow, J., & G. Aist. *The Sounds of Silence: Towards Automated Evaluation of Student Learning in a Reading Tutor that Listens.* In Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97). 1997. p. 355-361 Providence, RI: American Association for Artificial Intelligence.

Mostow, J., & Aist, G. (2001). Evaluating tutors that listen: An overview of Project LISTEN. In P. Feltovich (Ed.), *Smart Machines in Education* (pp. 169-234). Menlo Park, CA: MIT/AAAI Press.

Mostow, J., G. Aist, J. Beck, R. Chalasani, A. Cuneo, P. Jia, and K. Kadaru. *A La Recherche du Temps Perdu, or As Time Goes By: Where does the time go in a Reading Tutor that listens?* In Proceedings of the Sixth International Conference on Intelligent Tutoring Systems (ITS'2002). 2002. p. 320-329 Biarritz, France: Springer.

Mostow, J., Aist, G., Bey, J., Burkhead, P., Cuneo, A., Junker, B., Rossbach, S., Tobin, B., Valeri, J., & Wilson, S. (2002, June 27-30). *Independent practice versus computer-guided oral reading: Equal-time comparison of sustained silent reading to an automated reading tutor that listens.* Presented at the Ninth Annual Meeting of the Society for the Scientific Study of Reading. 2002. Chicago, Illinois.

Mostow, J., J. Beck, R. Chalasani, A. Cuneo, and P. Jia. *Viewing and Analyzing Multimodal Human-computer Tutorial Dialogue: A Database Approach.* In Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces (ICMI 2002). 2002. p. 129-134 Pittsburgh, PA: IEEE.

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

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what Students Know: The Science and Design of Educational Assessment*. National Research Council, Washington, D.C.: National Academy Press.

Sison, J., and J. E. Beck. *Using Speech Data to Assess Reading Proficiency.* Presented at the Computer Assisted Language Instruction Consortium conference in June. 2004. Pittsburgh, PA.

Webb, G. I., & Kuzmycz, M. (1998). *Evaluation of data aging: A technique for discounting old data during student modeling*. Paper presented at the Interantional Conference on Intelligent Tutoring Systems, San Antonion, TX.

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

Huang, X.D., F.vAlleva, H.W. Hon, M.Y. Hwang, K.F. Lee, and R. Rosenfeld. The SPHINX-II Speech Recognition System: An Overview. *Computer, Speech and Language,* 1993. 2: p. 137-148.