Can a Reading Tutor that Listens use Inter-word Latency to Assess a Student's Reading Ability?

Peng Jia[1], Joseph E. Beck[2], and Jack Mostow[2]

Project LISTEN, [1]Center for Automated Learning and Discovery and [2]Robotics Institute
Carnegie Mellon University, Pittsburgh, PA 15213-3890 pengj@cs.cmu.edu

Abstract. This paper describes our use of inter-word latency, the delay before a student speaks a word in the course of reading a sentence, to assess oral reading automatically. The context of our study is a Reading Tutor that uses automated speech recognition to listen to children read aloud. Using the data from 58 students from grades 1 through 4, we used inter-word latency to predict scores on external, individually administered, paper-based tests. Correlation between predicted and actual test scores exceeded 0.7 for fluency, word attack, word identification, word comprehension, and passage comprehension. Compared with paper-based tests, this evaluation method is much cheaper, based on computer-guided oral reading recorded in the course of regular tutor use, and invisible to students. It has the potential to provide continuous assessment of student progress, both to report to teachers and to guide the Reading Tutor's own tutoring.

Keywords: assessment, psychometrics , latency, Woodcock Reading Mastery Tests, Intelligent Tutoring Systems, Reading Tutor, fluency

1 Introduction and Motivation

Project LISTEN's Reading Tutor [1] is an ITS that listens to children read aloud and provides feedback to students. Project LISTEN currently uses the Woodcock Reading Mastery Tests, a battery of widely used reading performance evaluation tests, to evaluate students' reading proficiency improvements. Researchers have already validated the content of these tests and have gathered psychometric data about these paper-based tests' reliability and validity for making predictions. Validated instruments provide a check on ITS researchers' claims, as gains made within the tutor may not translate to contexts outside of the tutor.

However, there are several drawbacks to using such instruments:

1. The tests take time to administer. Schools are not always willing to have their students spend time taking tests that could be spent on educational activities.
2. Many tests, particularly those for young children, must be individually administered and require training to administer properly.
3. If students are absent when the test is administered, and the goal of the study is to relate gains on an instrument to a particular intervention, the students would be lost from the pool of subjects.
4. Assessment can be measured at the beginning and the end of the study, but it can be difficult to infer a student's progress in the interim.
5. Tests must be purchased from the publisher, which is an additional expense.

Our short-term goal is to find a method to assess student performance that does not suffer from, or at least not suffer as badly, the above problems. Once our method is validated against established reading proficiency assessment methods, it would not be necessary to continue to test students each year with the paper-based tests. We would be able to help school instructors' teaching by providing them a set of reading proficiency assessments for each student.

Currently, the Reading Tutor does not have a strong student model and there are some difficulties in evaluating student reading performance within the tutor. If an assessment method can be automated to provide the Reading

Tutor a useful measure for tracking student progress, then the tutor could add this information to its own student model. This helps realize our long-term goal of improving the Reading Tutor.

We use inter-word latency in this paper as the automated measure of student reading ability. Our research is an extension of prior work on the Reading Tutor investigating inter-word latency [2], which is the time that elapses between words that a student reads.

[2] reported significant difference in average latencies for 36 stop words (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) vs. all words that are not in this 36 words list for 8 low-reading third graders. This work also showed that latency reflects improved reading performance since latency decreases significantly with Reading Tutor use. Therefore, proposing latency as a reading performance measure is a natural conclusion. However, [2] did not relate latency to established performance measures. Thus, we are not certain whether changes in latency are consistent with verifiable claims about the student's reading proficiency.

Previous research has studied the relation between time-based measures of student reading to other reading assessment measures. [3] used data from 134 fourth graders and found a significant negative correlation between the time to read a word and reading comprehension. However, their time measures are for isolated words, whereas our inter-word latency measure applies to connected text.

In this paper, we relate inter-word latency to a measure of reading fluency and to the Woodcock Reading Mastery Tests [4]. Determining if there is a relation between the inter-word latency measure and the external paper-based tests can help justify using latency to make decisions within the Reading Tutor.

2 Approach

In this section, we first give the definition of inter-word latency and describe the subset of inter-word latencies we chose to use in this study. We then briefly introduce paper-based fluency tests as well as Woodcock Reading Mastery Tests. Finally, we describe the dataset we used in this study.

2.1 Description 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.

A student could correctly read a word; he could misread it; or he could omit it. Table 1 provides an example of these concepts. If the actual sentence text is: "It was the worst quake ever" and the student read "*it was the were…ever*," then the word "quake" was skipped by the student; the word "worst" was misread while all other words were read correctly.

Actual sentence text	Speech recognizer output for each word	Start time (ms)	End time (ms)	Latency (ms)
It	IT	0	360	N/A
was	WAS	480	610	120
the	THE	650	860	40
worst	WERE	1040	1140	N/A
quake				N/A
ever	EVER	1570	1640	N/A

Table 1. Latency computation

The inter-word latency for a word w_i, the i^{th} word in the actual sentence to be read, is defined as follows:

i. If w_i was read correctly starting at time $t_{i,start}$

ii. And if w_{i-1} was read (either correctly or misread) ending at time $t_{i-1,end}$

iii. Then, the inter-word latency for word w_i is $t_{i,start} - t_{i-1,end}$.

Thus, inter-word latency is only defined for words that are correctly read and are immediately preceded by a word that was not omitted by the student. Therefore, the first word of any sentence will never have an inter-word latency. This is the case for the word "it" in Table 1. There is no latency for word "worst" since it was not read correctly. The word "quake" does not have a latency measure either because it was omitted by the student. This omission also causes the word "ever" to have no inter-word latency by the second condition in the inter-word latency definition. The other words in the sentence have latency measures as shown in the last column in Table 1. For example, for the word "the," the student started to read the word at 650 ms, and finished reading the word "was" at 610 ms. Therefore, the latency is 650-610 = 40 ms.

2.2 Which latencies to consider?

A student could make several attempts at reading a sentence. For this study, we only consider the first attempt since it usually reflects the student's true reading skill better because, in the following attempts, he might just repeat what he had said before which would artificially shorten the latencies.

A student might encounter the same word several times when he used the Reading Tutor. To assess student's initial reading proficiency, for each student, we computed the latency for the first time a student encountered a word, and define this as *initial latency.* We excluded data gathered during a 14-day period after each student started using the Reading Tutor since these data are conflated with students learning of how to use the Reading Tutor.

To assess each student's reading proficiency at the end of the school year, we defined each word's *final latency* as follows: we computed the latency only his *first* attempt at reading a word on the *last* day that he encountered that word. Subsequent exposures to the same word on a given day might artificially shorten the latency due to effects of recent practice. To avoid this confound, we do not consider those encounters. A further restriction on final latencies was it must occur in a story that the student had never read before. Prior experience with the Reading Tutor suggested that there were stories that were favorites of students, and students had memorized these stories through repeated exposure. Counting this memorization would bias the latency measure since it is supposed to be measuring reading.

We used only those words that had both an initial and a final latency. This allowed us to compare initial latencies with final latencies without worrying about controlling for word difficulty. Another reason was to replicate the study done in [2] that also used paired latencies.

[2] reported significant difference in average latencies of "easy" and "hard" words, namely 36 stop words and words not in this 36 word list. Following this finding, we separated words in our paper using a slightly different approach. The Dolch list [5] is often used in reading proficiency studies [6, 7]. This list has 220 very frequent words used by children's books and covers all the 36 stop words used in [2]. Since these words are words that "glue" a sentence's content together, a student must recognize them quickly so as not to impede the comprehension of the sentence [8]. We thus assumed that average inter-word latencies for Dolch words would be lower than that for non-Dolch words and we separated and compared these two types of words.

The Reading Tutor has spelling activities where the student has to spell (for example) "CAT." In this case, the Reading Tutor would count "C" and "A" and "T" as encountered "words." Since these single character "words" are not real words, we removed them from the dataset. Furthermore, words that are expressed in numerical format (e.g. 33, 1999….) were also removed from the pool of data.

2.3 Description of Fluency and Woodcock Reading Mastery Tests

For external evaluation, we use pre- and post-test scores of the fluency and the Woodcock Reading Mastery Tests. The fluency testing was done by having students read three passages that were at their grade level and then calculating the number of words the student read correctly in 60 seconds (taking the median of the scores on the 3 passages). The Woodcock Reading Mastery Tests are a battery of tests that produces an overall score for

reading proficiency and four subscores for deciphering unfamiliar words (word attack), identifying single words (word identification) and understanding words and passages (word comprehension and passage comprehension).

All of these tests were individually administered and scored by hand. Pretests were given in September/October and post-tests were given in April/May. Our approach is to relate students' pre-test scores with their initial latency and post-test scores with their final latency.

2.4 Description of 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. We parsed and loaded these log data into the database [9]. Currently, the database includes data for 58 students who used the Reading Tutor for the entire academic year of 2000-2001 and for whom we had pre- and post-test scores for the four subtests of the Woodcock Reading Mastery Tests and fluency scores. These students were in grades 1 through 4 (i.e. 6 to 9 year-olds), with 25 students in grade 1, 13 students in grade 2, 11 students in grade 3, and 9 students in grade 4. The number of boys and girls in each grade distribute roughly evenly for all grades (except for grade 3, which had 4 girls vs. 7 boys).

In total, there are 31,437 pairs of initial and final latencies. Each student averaged 524 pairs (minimum=114, maximum=1737, median=411 pairs). Among these latency pairs, there are about 26% (minimum=10%, maximum=39%, median=27%) are for words in the Dolch list for each student.

3 Results

We now discuss relating our latency measure with the paper-based tests.

3.1 Reliability and Statistical Properties of the Latency Measure

Latency for each word is very noisy and thus cannot be used to make meaningful predictions. Considering the average latencies of all words (or a large pool of words) that each student read smoothes out the noise and results in a more useful measure. Therefore, we computed average initial and final latencies for each student and used these aggregated results instead of latencies for individual words. Table 2 summarizes how students' average latencies varied. A paired T-test shows that, for the 58 students, average final latencies were significantly shorter than average 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 ($P < 0.002$). This also confirmed [2] that, on average, latency could reflect the difficulties of words.

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

Table 2. Descriptive statistics for average initial and final latencies of all 58 students

We also studied the reliability of latency by using the test-retest methodology. The students' average initial latencies correlate 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 gives a correlation of 0.79 for initial latencies and 0.64 for final latencies. Using the Spearman-Brown prophecy formula correction [10] gives a reliability of 0.88 and 0.78 for average initial latencies and average final latencies, respectively. Therefore, latency is a reliable measure.

3.2 Construct Validity

The inter-word latency, the fluency tests, and the Woodcock Reading Mastery Tests all 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 Reading Tutor's speech recognizer's accuracy, and considers only words that the tutor recorded as being correctly read (and pronounced). In fluency tests, passages read by the students are at their grade level. The latency measures do not have any guarantees about the type of words that are included, but the Reading Tutor does attempt [11]to give students passages that are at their level of reading ability. Thus, both the fluency and latency measure seem to measure the same underlying construct: "how fast the student reads."

How our latency measure maps to the Woodcock Reading Mastery Tests is less clear. The Woodcock Reading Mastery Tests break reading into component skills and assess each skill individually. The latency measure takes one facet of reading—how long a student delays before reading a word—which we then use as a general measure of reading ability. Thus, there is considerably less overlap between the content of the Woodcock Reading Mastery Tests and our latency measure than between the latency and fluency measures. However, [12] reported a 0.9 correlation between reading comprehension and fluency. This is not surprising, since automaticity in recognizing words would both speed up reading and provide more working memory resources for comprehension (rather than spending those resources decoding individual words). Therefore, while our latency measure does not directly measure comprehension, it is plausible that it measures a closely related construct.

3.3 Statistical Validation

Does latency correlate with fluency? Figure 1 shows the relationship between each student' fluency pretest score and his average initial latency on non-Dolch words. The line is a lowess curve (constructed via SPSS's built in function) generated by fitting 75% of the data. The graph shows no linear relationship between the two measures but it indicates we might get a linear relation between fluency and latency by taking the inverse of average latency for each student.

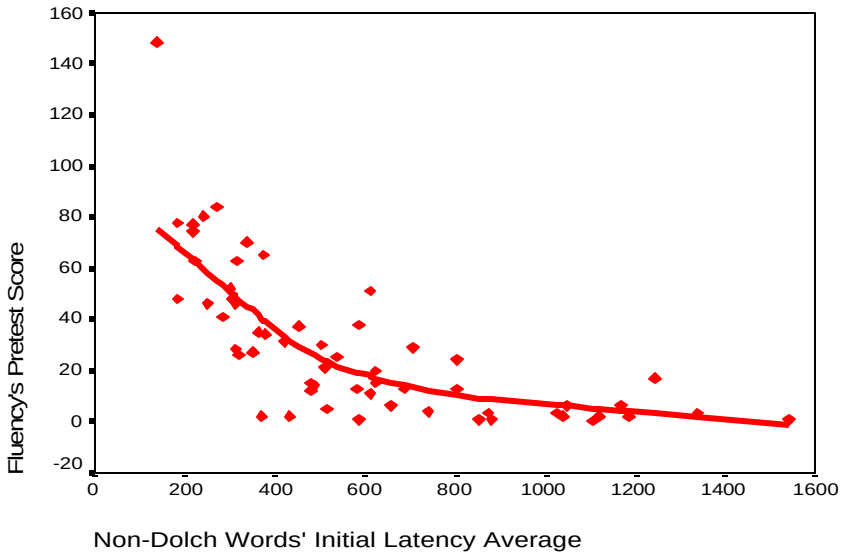

Figure 1. Non-Dolch words' average initial latency vs. fluency pretest for each student

We constructed a regression model using the inverse of the latency for non-Dolch words. Figure 2 shows the predictions of this regression model plotted against each student's fluency pretest score. In this Figure, each point is labeled by the student's grade. This relationship shown in Figure 2 is linear, with a correlation of 0.86. The regression equation is *fluency* = 18227 / *latency* – 13 and has an adjusted R^2 of 0.74. Visual inspection suggests that the regression does not fit very well for first graders who did poorly on the fluency pretest.

One concern is that students' fluency correlates with their grade level. Therefore, pooling students together may inflate the actual relationship between latency and fluency. Controlling for grade gives a partial correlation of 0.74 between predicted and actual latency. Although constructing separate regression equations for each grade is the best way to model the data, our database does not yet have enough data for four models.

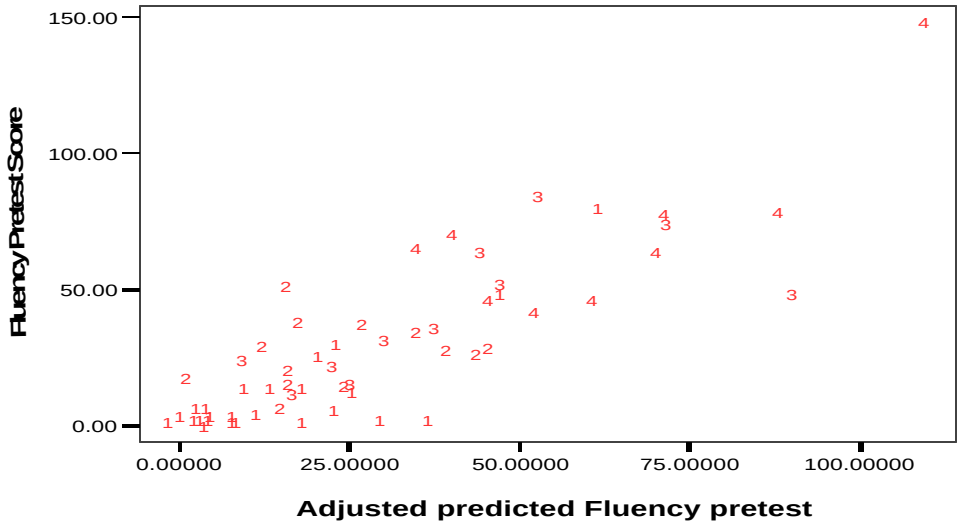

Figure 2. Actual fluency vs. predicted fluency

Figure 3 shows that there is not a linear relation between our inverse latency measure and the word identification scores. Specifically, for a high inverse latency (i.e. for students who read quickly) there was a point past which further speedups would not correspond to improvement on the Woodcock Reading Mastery Tests (this trend was true for the other Woodcock Reading Mastery Tests scores). To account for this, we take the logarithm of the inverse average latencies. Figure 4 shows the results of a regression model built using this logarithm transformation. This model has an adjusted R^2 of 0.69.

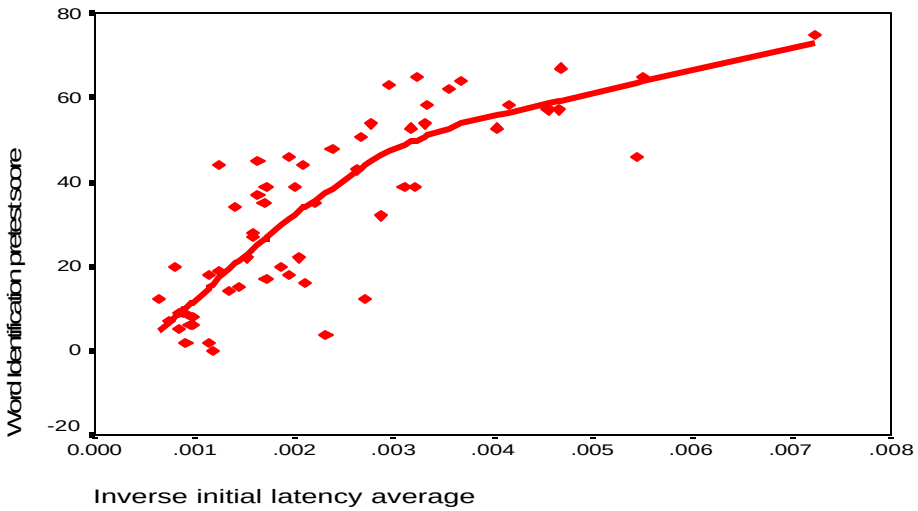

Figure 3. Word identification pretest scores vs. inverse initial latency averages

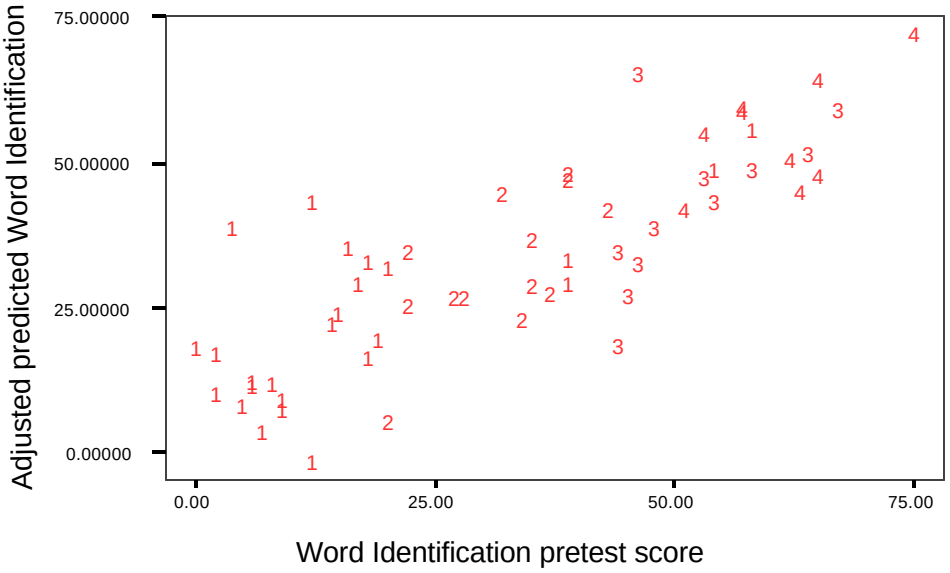

Figure 4. Actual vs. predicted word identification pretest score

We correlated students' average initial latencies with pre-test scores and average final latencies with student post-test scores and found:

1. Latency correlates with all of the Woodcock Reading Mastery Tests.
2. The correlations between pre-test scores and average initial latencies are higher than that for post-test scores and average final latencies.
3. The correlations between fluency and latency for both pre- and post tests scores for Dolch words' are, on average, lower than that for Non-Dolch words' latencies.

Table 3 shows these results in more detail:

	Pretests		**Posttests**	
	Average Initial Latencies		**Average Final latencies**	
	Non Dolch	**Dolch**	**Non Dolch**	**Dolch**
Word identification	0.83	0.73	0.64	0.55
Word attack	0.72	0.64	0.53	0.45
Word comprehension	0.80	0.71	0.61	0.51
Passage comprehension	0.82	0.74	0.63	0.53
Fluency	0.86	0.74	0.60	0.60

Table 3. Correlations between paper-based test scores and latencies (all significant at p<0.01)

For comparison, Table 4 shows how well the latency, fluency, and Woodcock tests inter-correlate. This provides some intuitions about how well our automated measures could potentially do. For example, the Woodcock comprehension tests correlate at 0.94 and 0.95 with the word identification subtest. This suggests that our automated measure could do a substantially better job at predicting this subtest than it does.

	Fluency	Word attack	Word identification	Word comprehension	Passage comprehension	Latency
Fluency	-	0.74	0.86	0.80	0.81	0.86
Word attack	0.74	-	0.87	0.83	0.85	0.72
Word ID	0.86	0.87	-	0.95	0.94	0.83
Word comprehension	0.80	0.83	0.95	-	0.92	0.80
Passage comprehension	0.81	0.85	0.94	0.92	-	0.82
Latency	0.86	0.72	0.83	0.80	0.82	-

Table 4. Inter-test correlations for fluency, Woodcock, and latency for all 58 students

4 Conclusion and Future Work

We tested latency's validity as an automated assessment measure by correlating it with paper-based test scores. The significant correlation between fluency and latency suggests that we might be able to avoid paper-based fluency tests and still provide teachers with a similar measure by using inter-word latencies. Through this first but very encouraging step, we see the potential of exploiting the rich student-tutor interaction data for automated internal assessment.

Latency is one of our first attempts at using an automated measure to evaluate students. We have different grain-size data from millions of student-tutor interactions that might contain other useful information that can be used to improve the Reading Tutor, such as the number of help requests made by students. The nature of machine-collected data (e.g. precise timestamps of events, ability to record unsupervised student activities) might help us find better ways to evaluate student's reading performance.

We must be cautious with replacing Woodcock Reading Mastery Tests by latency, however, as latency may be measuring a different construct than the Woodcock Reading Mastery Tests. Thus, while such low-level data may be useful for evaluating student performance while using the Reading Tutor, attempting to predict Woodcock Reading Mastery Tests scores needs further study. Also, different measures will have different ceilings and floors beyond which they lose predictive power; the precise limit depends on the measure predicted. As seen in Figure 3, students who read faster probably are better at identifying words. However, at some point this relationship will start to break down. Determining where this breakdown occurs is critical.

In the future, the following consideration might help us in better modeling student latency:

1. Lengthy latencies might occur in two kinds of situations:
   i. Situations where there were interaction problems between the student and the Reading Tutor, such as not agreeing on what part of the sentence to read.
   ii. Situations where the student was in fact 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. Project LISTEN administered paper-tests to assess student's fluency 4 times per academic year. We only used the first and the last tests' in this paper. Instead of only relating initial and final latencies to pretest and posttest scores, we can correlate four tests fluency tests, (pre-test, two interim tests and post-test) with latency data from the specific months when the paper-based tests were administered. This way, we can study the relations between latency and fluency in a finer way. We could also gain some knowledge about how latencies tend to change during the course of the year.

Our initial examination of the data has left us with several unanswered questions:

1. Why are the correlations of latency with post-test scores consistently lower than those between latency and pre-test scores?

2. Why do correlations between latency and post-test Woodcock Reading Mastery Tests scores differ significantly within subpopulations? For modeling pretests, we do about as well for boys (N=31) as for girls (N=27) with correlations averaging 0.79 for girls and 0.82 for boys (P=0.15). However, for modeling posttests correlations for girls average 0.75 while boys average 0.50. This difference is significant at P<0.001.

   This might be an effect of small sample (N=31 for boys and N=27 for girls), or could indicate systematic differences in how children used the tutor. We cannot be certain about this due to small sample size. We will have a bigger dataset to test because our database is still being populated with students who used the Reading Tutor in the 2000-2001 academic year. If the result of future experiments on a larger dataset still shows this population difference, we may need to model each subpopulation separately or determine what types of behaviors in the tutor cause us to better estimate one group than another.

5 Acknowledgments

This work was supported in part by the National Science Foundation under Grant No. 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 thank other members of Project LISTEN who contributed to this work, especially Andrew Cuneo for getting the tutor interactions into database form, Susan Rossbach for supervising the administration of the Woodcock Reading Mastery Tests and fluency tests, and Brian Tobin for getting the test data into easily analyzable form. We also thank the students and educators at the schools where the Reading Tutor records data.

References

1. Mostow, J. and Aist, G., *Evaluating Tutors that Listen: An Overview of Project LISTEN*, in *Smart Machines in Education: The Coming Revolution in Educational Technology*, K.D.F.a.P.J. Feltovich, Editor. 2001, AAAI Press/ The MIT Press: Menlo Park, Carlifornia. p. 169-234.

2. Mostow, J. and Aist, G. *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.

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

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

5. Dolch, E.W., *Dolch Word List*. 1936.

6. *Usage of Dolch word list* *http://www.senter.co.uk/Information/dolch%20list%20and%20other%20vocabulary%20lists%20info.htm*.

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

8. May, F.B., *Reading as Communication: To Help Children Write and Read*. 5 ed. 1998, Upper SaddleRiver, NJ: Prentice Hall. 557.

9. Mostow, J., Aist, G., Beck, J., Chalasani, R., Cuneo, A., Jia, P., and Kadaru, K. *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. Biarritz, France: Springer.

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

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

12. Deno, S.L., *Curriculum-Based Measurement: The emerging alternative.* Exceptional Children, 1985. **52**(3): p. 219-232.