Performance of Automated Scoring for Children's Oral Reading

Ryan Downey, David Rubin, Jian Cheng, Jared Bernstein
Pearson Knowledge Technologies
299 S. California Ave.
Palo Alto, California 94306

Ryan.Downey@Pearson.com

Abstract

For adult readers, an automated system can produce oral reading fluency (ORF) scores (e.g., words read correctly per minute) that are consistent with scores provided by human evaluators (Balogh et al., 2005, and in press). Balogh's work on NAAL materials used passage-specific data to optimize statistical language models and scoring performance. The current study investigates whether or not an automated system can produce scores for young children's reading that are consistent with human scores. A novel aspect of the present study is that text-independent rule-based language models were employed (Cheng and Townshend, 2009) to score reading passages that the system had never seen before. Oral reading performances were collected over cell phones from 1st, 2nd, and 3rd grade children (n = 95) in a classroom environment. Readings were scored 1) *in situ* by teachers in the classroom, 2) later by expert scorers, and 3) by an automated system. Statistical analyses provide evidence that machine Words Correct scores correlate well with scores provided by teachers and expert scorers, with all (Pearson's correlation coefficient) *r*'s > 0.98 at the individual response level, and all *r*'s > 0.99 at the "test" level (i.e., median scores out of 3).

1 Introduction

Oral reading fluency (ORF), defined as "the ability to read a text quickly, accurately, and with proper expression" (National Reading Panel, 2000; p. 3.5), is a reflection of readers' decoding ability. Skilled readers can recognize words effortlessly (Rasinski and Hoffman, 2003), due to "automaticity" of processing (LaBerge and Samuels, 1974) whereby a reader's attention is no longer focused on "lower level" processing (e.g., letter to phoneme correspondence, word identification, etc.). Instead, attention can be devoted to "higher level" functions such as comprehension and expression (LaBerge and Samuels, 1974). As a means of assessing general reading ability, oral reading fluency performance is also a predictor of student success in academic areas such as reading and math (e.g., Crawford, Tindal, and Stieber, 2001). Oral reading fluency is one of the key basic skills identified in the Reading First initiative used to satisfy the standards of the No Child Left Behind Act (NCLB, 2001).

Although oral reading fluency is comprised of several abilities, due to practical constraints the most commonly reported reflection of oral reading fluency is reading rate, specifically, the words read correctly per minute (WCPM). Typically, ORF performance is measured by a classroom teacher who sits alongside a student, marking and annotating – in real time – the student's reading on a sheet of paper containing the passage to be read. Classroom testing is time-consuming and requires a teacher's full attention. In practice, teaching time is often sacrificed to "testing time" to satisfy local and federal reporting standards (e.g., NCLB). ORF scoring guidelines are specific to particular publishers; teachers must undergo training to become familiar with these guidelines, and cost, availability, and quality of training varies. Finally, despite good-faith attempts to score accurately, teachers may impose errors and inconsistencies in

Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications, pages 46–55,
Portland, Oregon, 24 June 2011.

scoring ORF performances due to unavoidable factors such as classroom distractions, varying experience with different accents/dialects, varying experience with scoring conventions, and differences in training, among others.

To address the need for a rapid and reliable way to assess oral reading fluency, a growing body of research has supported the use of automated approaches. Beginning with work by Bernstein et al. (1990) and Mostow et al. (1994), prototype systems for automatic measurement of basic components of reading have appeared. Recent projects have addressed finer event classification in reading aloud (Black, Tepperman, Lee, Price, and Narayanan, 2007), and word level reading (Tepperman et al., 2007), among others. Research has increasingly focused on systems to score passage-level reading performances (e.g., Balogh et al., 2005; Zechner, Sabatini, and Chen, 2009; Cheng and Townshend, 2009). Eskenazi (2009) presents a general historical perspective on speech processing applications in language learning, including reading.

The present automated ORF assessment was developed to deliver and score tests of oral reading fluency, allowing teachers to spend less time testing and more time teaching, while at the same time improving score consistency across time and location. Automated ORF tests are initiated by a click in a web-based class roster. Once a test is initiated, a call is placed to a local phone number and the test begins when the phone is answered. Instructions presented through the handset direct the student to read passages out loud into the cell phone, and these readings are sent to the automated ORF system for processing and scoring.

2 Present Study

The scoring models used by the automated ORF test (see **Method** below) were originally developed based on adult readings, and then optimized on large sets of data collected from students reading passages produced by AIMSweb, a publisher of Reading Curriculum-Based Measurement (R-CBM) oral reading fluency passages (www.aimsweb.com). AIMSweb passages are leveled and normed across large samples of students. Previous validation studies found that when the system was optimized using data from students reading AIMSweb passages, machine scores correlated with trained human expert score with r = 0.95 to 0.98, depending on the grade level of the student readers.

The primary question that the present studies attempt to answer is whether the automated scoring system can score newly inserted content – in this case, ORF passages offered by Sopris called "Dynamic Indicators of Basic Early Literacy Skills", or DIBELS (www.dibels.com) – accurately and at a high level of reliability. This is an evaluation of text-independent Rule Based Language Models (RBLMs) that were developed with training data from *other* readers performing on *other* passages and then applied to the new passages.

A secondary question of interest involves how different types of scorers may assign Words Correct scores differently. Two groups of human scorers were recruited: 1) teachers who were recently trained in DIBELS scoring methods who would perform scoring in the classroom, and 2) expert scorers with the ability to score reading recordings carefully and at their convenience, without classroom distractions. Answering the first part of the question involves comparing machine Words Correct scores to human scores when teachers make ratings in the classroom environment as the student reads into the phone. This analysis reveals if the machine and teachers produce systematically different scores when testing is performed in a "live" classroom with the typical attentional demands placed on a teacher scoring an ORF passage. Answering the second part of the question involves comparing machine Words Correct scores to a "consensus", or median Words Correct value, from expert scorers. These three experts, with over 14 years of combined experience scoring DIBELS passages, listened to recordings of the same readings made in the classroom. Because the recordings were digitally preserved in a database, the expert scorers were able to replay any part(s) of the recordings to determine whether each word was read correctly. The benefit of being able to replay recordings is that such scores obtained are, in theory, closer to capturing the "truth" of the student's performance, unaffected by biases or distractions encountered by scorers performing a "live" rating.

2.1 Method

2.1.1 Rule Based Language Models

The scoring models used by the automated ORF system are RBLMs such as those described by Cheng and Townshend (2009). Such models outperform traditional n-gram language models (Cheng and Shen, 2010), in part by adding intuitively simple rules such as allowing a long silence as an alternative to a short pause after every word, leading to improvements in accuracy. Also, rules like those described by Cheng and Townshend (2009) consider much longer sequential dependencies. The basic idea for this kind of language model is that each passage gets a simple directed graph with a path from the first word to the last word. Different arcs are added to represent different common errors made by the readers, such as skipping, repeating, inserting, and substituting words. For each arc, a probability is assigned to represent the chance that the arc will be chosen. Knowledge of performance on other readings produces linguistic rules, such as *she* can substitute for *he*, a single noun can replace a plural noun, the reader may skip from any place to the end, etc. All the rules used in RBLMs can be classified into five broad groups:

1. skip/repeat rules
2. rules using part-of-speech (POS) tagging information
3. rules accommodating for insertion of partial words
4. general word level rules
5. hesitation and mouth noise rules

A detailed analysis of the role of rules in RBLMs was described in Cheng and Shen (2010).

The language rules are extrapolated from transcriptions of oral reading responses to passages using four base rules: any word substitutes for any word with a low probability; any word is inserted after any word with a low probability; any word is skipped with a low probability; any word is repeated immediately with a low probability. Following Cheng and Townshend (2009), the first two are the only rules that allow out-of-vocabulary words and their probabilities are fixed to the lowest level, so their arcs will never be traversed unless there is no other choice.

General language model rules for reading can be inferred from clustering traversals of the basic models and proposing further rules that can be applied to new reading passages and used to infer underlying knowledge about the reading. Arcs are added to represent commonly observed non-canonic readings. Further analysis of rule-firing details may provide diagnostic linguistic information about children's reading habits that can be reported and analyzed.

In the present automated scoring system, new passages are automatically tagged for part-of-speech (POS) using the Penn Tree Tagger (Marcus, Santorini, and Marcinkiewicz, 1993). POS tags allow specification of certain general rules based on linguistic properties, such as:

- NN (noun, singular or mass) can become NNS (noun, plural);
- VBZ (verb, 3rd person singular present) can become VBP (verb, non-3rd person singular present); and so on.

These patterns occur quite frequently in real responses and can therefore be accounted for by rules. Sentence, clause, and end-of-line boundaries are tagged manually. Marked up passages are then inserted into the ORF scoring system, providing data regarding places in the reading that may result in pauses, hesitations, corrections, etc. If the expected response to a reading passage is highly constrained, the system can verify the occurrence of the correct lexical content in the correct sequence. It is expected that the system, using previously trained data coupled with the RBLMs from the newly inserted passages, will be able to produce Words Correct scores with high accuracy (i.e., consistent with human Words Correct scores).

Here, we make a final note on the use of Words Correct instead of words correct per minute (WCPM), when WCPM is the most common measure for quantifying oral reading performance. The automated system presents students with a 60-second recording window to read each passage, but it calculates a truer WCPM by trimming leading and trailing silence. Human scorers simply reported the number of words correct, on the assumption that the reading time is the recording window duration. Thus, Words Correct scores are the appropriate comparison values, with a fixed 60-second nominal reading time.

2.1.2 Participants

A total of 95 students were recruited from the San Jose Unified School District in San Jose,

California. The students were 20 first graders, 20 second graders, and 55 third graders, all enrolled in a summer school program. Students with known speech disorders were included in the study, as was one student with a hearing impairment. Roughly half of the participants were male and half were female. A number of English Language Learners are known to have been included in the sample, though language status was not recorded as a variable for this study. It is not known whether any of the students had been diagnosed with reading disabilities.

Four Teachers were trained to administer and score DIBELS ORF passages by an official DIBELS trainer, over the course of a two day training session. All Teachers were reading experts or teachers with experience in reading education. They were trained to navigate a web application that triggers delivery of tests over cell phones under classroom testing conditions. Evaluator qualifications are summarized in Table 1.

Evaluator	Highest degree, or relevant certification	Years assessing reading
Teacher 1	MA Education	8
Teacher 2	MA Education	7
Teacher 3	Reading Credential	15
Teacher 4	BA Education	12
Expert 1	MS, Statistics	5
Expert 2	EdS, Education	2
Expert 3	MA Education	20

Table 1. Evaluator qualifications

2.1.3 Procedure

First, nine passages – three for each of the three grades, presented together in a single test – were drawn from the DIBELS Benchmark test materials. Each DIBELS passage was tagged for parts of speech and formatting (e.g., line breaks) and inserted into the automated scoring system. Rule-based language models were produced for each passage.

During data collection, each student read the grade-appropriate DIBELS Benchmark test (3 passages) into a cellular telephone in the classroom. With three passages per student, this process yielded 285 individual reading performances.

Once a test was initiated, Teachers allowed the test to run independently and scored manually alongside the student reading into the phone. According to standard DIBELS scoring conventions, the students were allowed to read each passage for one minute. Teachers calculated and recorded the Words Correct score on a worksheet for each passage. Teachers returned the annotated score sheets for analysis.

Later, three Expert scorers logged in to a web-based interface via the Internet, where they listened to the digitized recordings of the readings. All three Expert scorers had extensive experience with DIBELS rating. One Expert was the DIBELS trainer who provided the DIBELS training to the Teachers for this study. Experts scored students' performance manually using score sheets with the instruction to use standard DIBELS scoring conventions. Each Expert entered a Words Correct score for each passage using the web interface, and the score sheets were returned for analysis.

2.1.4 Automated scoring

Incoming spoken responses were digitally recorded and sent to a speech processing system that is optimized for both native and non-native speech. Recognition was performed by an HMM-based recognizer built using the HTK toolkit (Young, et al., 2000). Acoustic models, pronunciation dictionaries, and expected-response networks were developed in-house using data from previous training studies involving many thousands of responses. The words, pauses, syllables, phones, and even some subphonemic events can be located in the recorded signal, and "words recognized" are compared with "words expected" to produce a recognized response and word count.

The acoustic models for the speech recognizer were developed using data from a diverse sample of non-native speakers of English. In addition, recordings from 57 first-grade children were used to optimize the automated scoring system to accommodate for characteristics specific to young children's voices and speech patterns. These participants produced 136 usable, individual reading samples. These samples were each rated by two expert human raters. Using this final training set, the scoring models were refined to the point that the correlation between human and machine scoring was 0.97 for WCPM.

2.1.5 Human scoring

During data preparation, it was noted that many of the teacher scores were several words longer than would be expected based on the machine scores. Further investigation revealed that teachers would occasionally continue scoring after the one minute point at which the system stopped recording a passage, perhaps because they hadn't heard the notification that the reading was complete. A total of 31 out of 285 instances (~10.8%) were found where teachers continued scoring for more than 3 words beyond the 1 minute recording window, leading to artificially inflated Teacher scores. This artifact of the testing apparatus/environment warranted making a careful correction, whereby all Teacher scores were adjusted to account for what the machine "heard". That is, words and errors which Teachers scored after the automated system stopped recording (i.e., to which the automated system did not have access) were subtracted from the original Teacher Words Correct scores. All Teacher Words Correct scores reported hereafter are thus "corrected".

For purposes of finding a "consensus" Expert score, the median of the 3 expert human scores for each passage was obtained and is referred to as $Expert_M$ in the following analyses.

Nine readings from eight separate students received no scores from teachers. Information was not provided by the teachers regarding why they failed to complete the scoring process for these readings. However, we made the following observations based on the teachers' marked-up scoring sheets. For three readings, the teacher's final score was blank when the student appeared to have skipped lines in the passage. It is possible that, despite recent scoring training, the teacher was uncertain how to score skipped lines in the readings and left the final score blank pending confirmation. For one reading, the teacher made a note that the system stopped recording well before one minute had expired because the child's reading was too quiet to be picked up, and the teacher did not record the final score on the score sheet. For one reading, the student did not hear the prompt to begin reading (confirmed by listening to the response recording) and therefore did not read the entire passage; the teacher did not enter a final score. For the four remaining readings, the teacher annotated the performance but did not write down the final score for unclear reasons.

We might have elected to fill in the teachers' final scores for these 9 readings prior to subjecting the data to analysis, especially in the cases where a teacher annotated the reading correctly on the score sheet but simply failed to record the final Words Correct score, perhaps due to oversight or not knowing how to handle unusual events (e.g., entire line of reading skipped). Excluding such readings from the analysis ensured that the teachers' scores reflected "their own" scoring – including any errors they might make – rather than our interpretation of what the Teachers *probably* would have written. In addition, to maintain the most conservative approach, whenever a single reading passage from a student lacked a teacher's score, all 3 of that student's readings were excluded. The decision to exclude all readings from students with only a single passage missing was made because relevant analyses reported below involve reporting median scores, and a median score for students lacking one or two passage scores would not be possible.[1] The final set of graded responses thus consisted of 261 responses from 87 students.[2]

2.2 Results

2.2.1 Score Group Comparisons

Words Correct scores from Teachers, $Expert_M$, and machine are displayed in Table 2. Repeated measures ANOVA with Scorer Type (machine, Teacher, $Expert_M$) as the repeated measure and Score Group as the between-subjects factor revealed a main effect of group for the 261

[1] The excluded 8 students produced 15 readings with all three (Machine, Teacher, Expert) scores. Machine scores vs. Teacher scores and Machine scores vs. $Expert_M$ scores for these 15 individual responses yielded correlations of (Pearson's) r = 0.9949 and 0.9956, respectively. Thus, excluding these responses from the larger dataset is unlikely to have significantly affected the overall results.

[2] In production, such a system would not commit these errors of omission. Readings that are unscorable for technical reasons can trigger a *"Median score not be calculated"* message and request a teacher to manually score a recording or re-administer the assessment. Also, anomalous performances where Words Correct on one passage is very different from Words Correct on the two other passages could return a message.

readings[3], $F(2, 520) = 9.912$, $p < .01$, $\omega^2 < .001$. Post-hoc pairwise comparisons [4] showed that Words Correct scores from Teachers were higher on average than both the machine and $Expert_M$ scores (higher by 1.559 and 0.923 words correct, respectively; both p's < .05). On the other hand, Machine and $Expert_M$ scores did not differ significantly from each other (diff = 0.636).

Although the ANOVA showed that the means In the above analysis were significantly different, the effect size was negligible: ω^2 was = .0002, indicating that Score Group by itself accounted for less than 1% of the overall variance in scores. These results indicate that, for all 261 passages, the $Expert_M$ and machine scores were statistically comparable (e.g., within 1 word correct of each other), while Teachers tended to assign slightly – but not meaningfully – higher scores, on average.

Next, comparisons were made using the median value of each student's three readings. Median Words Correct scores for the 87 individual students were subjected to repeated measures ANOVA with the same factor (Scorer Group). Teachers' Words Correct scores were again higher than $Expert_M$ scores (diff = 1.115) and Machine scores (diff = 0.851), but this was not statistically significant in the main analysis, $F(2, 172) = 3.11$, $p > .05$, $\omega^2 < .001$. Machine Words Correct scores were, on average, 0.264 words higher than $Expert_M$ scores, but this, too, was not statistically significant. These results support the previous comparisons, in that machine scores fall well within ~1 word correct of scores from careful experts, while teachers tended to give scores of about 1 word correct higher than both experts and machine.

2.2.2 Scorer performance

To compare reliability, the Pearson's Product Moment coefficient (r) was used to estimate the correlation between paired human and machine scores, and between pairs of human raters. Two types of analyses are reported. First, analyses of Words Correct scores were conducted across scorers. Next, analyses were conducted on the basis of the median Words Correct score for each student's readings (i.e., the median score across all three passages). This score reflects the "real-life" score of DIBELS ORF tests because the median score is the one that is ultimately reported according to DIBELS scoring/reporting conventions.

2.2.2.1. Intra-rater reliability

Each Teacher scored each reading once during the live grading; intra-rater reliability could thus not be estimated for the Teacher group. During Expert rating, a randomly selected 5% of the passages were presented again for rating to each scorer. Overall Expert intra-rater reliability was 0.9998, with intra-rater reliability scores for Expert 1, Expert 2, and Expert 3 at 0.9996, 1.0, and 1.0, respectively. These results indicate that Expert human scorers are extremely consistent when asked to provide Words Correct scores for reading passages when given the opportunity to listen to the passages at a careful, uninterrupted pace. The automated scoring system would produce the exact same score (reliability = 1.0) every time it scored the same recordings, making its reliability comparable.

	Words Correct	
Score Type	**261 readings Mean (*SD*)**	**87 students Mean (*SD*)**
Teacher	84.3 *(42.5)*	84.0 *(42.1)*
$Expert_M$	83.4 *(42.3)*	82.9 *(41.8)*
Machine	82.8 *(39.6)*	83.2 *(39.3)*

Table 2. Mean Words Correct for all readings and all students.

2.2.2.2. Inter-rater reliability

Pearson's r was used to estimate the inter-rater reliability. All three Experts scored all passages, whereas any particular Teacher scored only a subset of the passages; thus, the Teacher's score was used without consideration of which teacher provided the score. Inter-rater reliability results are summarized in Table 3.

[3] For both ANOVAs, uncorrected degrees of freedom are reported but reported F values are corrected using Huynh-Feldt estimates of sphericity.

[4] Using Bonferroni adjustment for multiple comparisons.

Reliability (N = 261)			
	Teacher	Expert 1	Expert 2
Expert 1	0.998		
Expert 2	0.999	0.999	
Expert 3	0.998	0.999	0.999

Table 3. Inter-rater reliability estimates for Expert scorers.

To provide a measure of a "consensus" expert score, the median score from all 3 Experts was derived for each passage, and then compared with the Teacher score. This comparison (Teacher vs. $Expert_M$) yielded a reliability of 0.999, $p < .01$. As shown in Table 3, all inter-rater reliability estimates are extremely high, indicating, in part, that teachers in the classroom produce scores that do not differ systematically from those given by careful experts.

2.2.3 Human-machine performance

Pearson's *r* was computed to estimate the correlations. The different scorer groups (i.e., $Expert_M$, Teacher, and Machine) provided similarly consistent scoring, as evidenced by high correlations between scores from the three groups. These correlations were maintained even when data were broken down into individual grades. Table 4 reveals correlations between Words Correct scores provided by all 3 scoring groups, for each grade individually, for all three grades combined, and finally for the median scores for all 87 students.

Grade level (N)	Machine ~ Teacher	Machine ~ $Expert_M$	Teacher ~ $Expert_M$
1st grade (54)	0.990	0.990	0.996
2nd grade (60)	0.990	0.991	0.999
3rd grade (147)	0.964	0.962	0.997
Grades 1-3 (261)	0.989	0.988	0.999
Only medians 87	0.994	0.994	0.999

Table 4. Correlations between Words Correct scores by Experts, Teachers, and machine.

All correlations are 0.96 or higher. Correlations are highest between Teacher and $Expert_M$, but correlations between machine and both human groups are consistently 0.96 or above. The relatively lower correlations between human and machine scores seen in the third grade data may be attributed in large part to two outliers noted in the Figures below. If these outliers are excluded from the analysis, both correlations between human and machine scores in the third grade rise to 0.985. (See below for discussion of these outliers.)

2.2.4.1. Teacher vs. Machine performance

Pearson's *r* was used to estimate the correlation between Teacher and Machine scores. First, the Teacher-generated Words Correct score and Machine-generated Words Correct scores were obtained for each of the 261 individual recordings, where the correlation was found to be $r = 0.989$, $p < .01$.

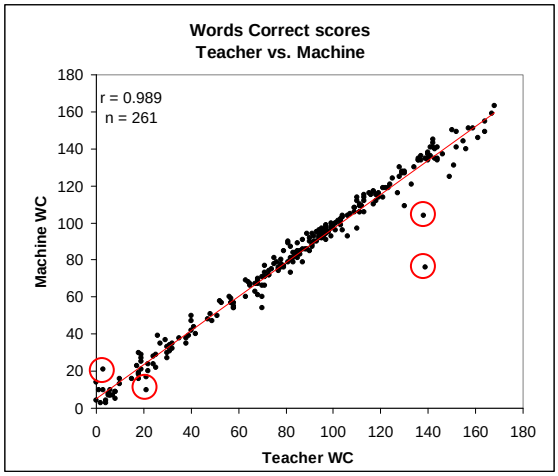

Figure 1. Words Correct (WC) scores from Teachers and Machine; response level (n = 261 responses)

Figure 1 shows a small number of outliers in the scatterplot (circled in red). One outlier (human = 3, machine = 21) came from a student whose low level of reading skill required him to sound out the letters as he read; machine scores were high for all 3 recordings from this reader. One outlier (human = 21, machine = 10) occurred because the reader had an unusually high pitched voice quality which posed a particular challenge to the recognizer. Two outliers (human = 141, machine = 76; human = 139, machine = 104) suffered from a similar recording quality-based issue whereby only some of the words were picked up by the system because the student read rapidly but quietly, making it difficult for the system to consistently pick up their voices. That is, for these calls the Teacher was close enough to hear the students' entire reading

but the machine picked up only some of the words due to distance from the telephone handset.[5]

Next, median Words Correct scores for each student were computed. Median scores derived from machine and Teachers correlated at 0.994, $p < .01$ for the 87 students. These scores are presented in Figure 2.

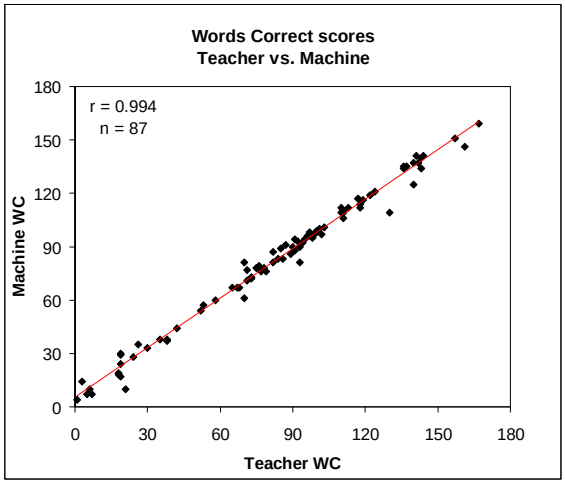

Figure 2. Words Correct (WC) scores from Teachers and Machine scoring at the reader level (n = 87).

Figure 2 shows that some of the outliers visible in the individual recording data disappear when the median score is computed for each student's reading performance, as would be expected.

2.2.4.2. Expert vs. Machine performance

The median of the 3 expert human scores for each passage ($Expert_M$) was compared to the Machine score. The correlation between machine-generated Words Correct scores and $Expert_M$-generated Words Correct scores was 0.988, $p < .01$, for the 261 individual readings, and 0.999, $p < .01$, for the median (student-level) scores. These results are displayed in Figure 3.

Figure 3 shows that two notable outliers present in the Teacher analysis were also present in the $Expert_M$ analysis. This may be due to the fact that while the recordings were of a low enough volume to present a challenge to the automated scoring system, they were of a sufficient quality for expert human scorers to "fill in the blanks" by listening repeatedly (e.g., with the ability to turn up the volume), and in some cases giving the student credit for a word spoken correctly even though they, the scorers, were not completely confident of having heard every portion of the word correctly. Though conjectural, it is reasonable to expect that the human listeners were able to interpolate the words in a "top down" fashion in a way that the machine was not.

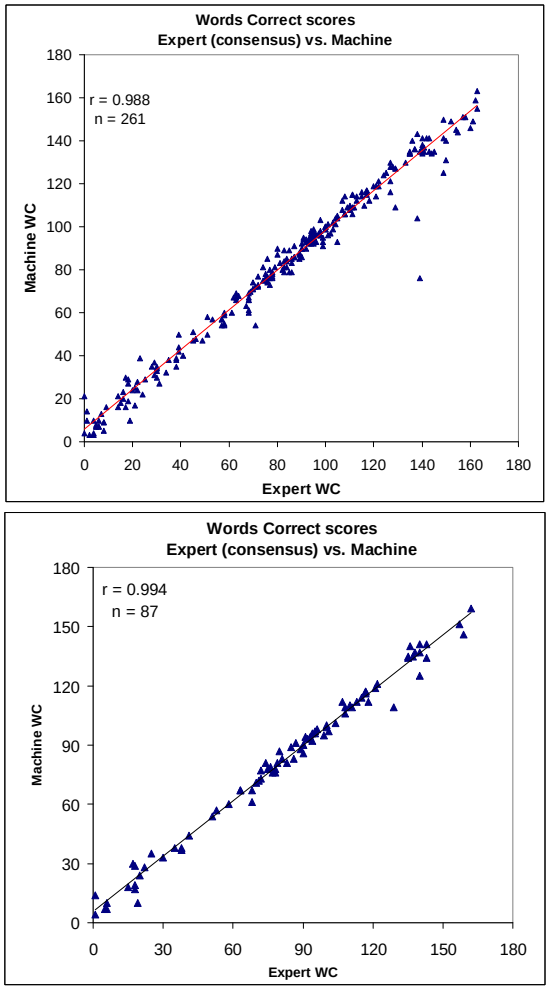

Figure 3. Words Correct (WC) Machine scores vs. Expert scores for all 261 individual responses (top) and for 87 students at test level (bottom).

2.2.5. Scoring Precision

It is reasonable to assume that careful expert scorers provide the closest possible representation of how a reading *should be scored,* particularly if the Expert score represents a "consensus" of expert opinions. Given the impracticality of having a team of experts score every passage read by a child

[5] In a production version, these recordings would return an instruction to re-administer the readings with better recording conditions or to score the recordings.

in the classroom, automated machine scoring might provide the preferred alternative if its scores can be shown to be consistent with expert scores. To explore the consistency between scores from Teachers and scores from the machine with scores provided by Experts, Teacher and Machine scores were compared against the median Expert score for each call using linear regression.

The standard error of the estimate (SE_E) for the two human groups was computed. The SE_E may be considered a measure of the accuracy of the predictions made for Teacher and Machine scores based on the (median, "consensus") Expert scores. Figure 4 below shows a scatterplot of the data, along with the R^2 and SE_E measures for both Teacher and machine scores based on Expert_M scores.

Scores from Teachers and Machine produce very similar regression lines and coefficients of determination (R^2 = 0.998 and 0.988 for Teachers and Machine, respectively). The figure also shows that, compared with the Machine scores, Teachers' scores approximate the predicted Expert_{Med} scores more closely (SE_E for Teachers = 1.80 vs. 4.25 for machine). This disparity appears to be driven by diverging scores at the upper and lower end of the distribution, as might be expected due to relatively smaller numbers of scores at the ends of the distribution.

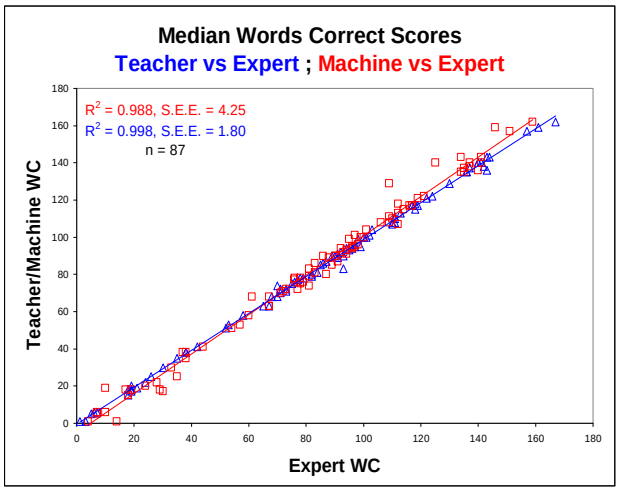

Figure 4. Median Words Correct scores from Machine (red squares) and Teachers (blue triangles) plotted against median Expert scores for 87 students. S.E.E. = Standard error of estimate.

3 Summary/Discussion

Correlations between human- and machine-based Words Correct scores were found to be above 0.95 for both individual reading passages and for median scores per student. The machine scoring was consistent with human scoring performed by teachers following along with the readings in real time (r = 0.989), and was also consistent with human scoring when performed by careful expert scorers who had the ability to listen to recorded renditions repeatedly (r = 0.988). Correlations were consistent with those between expert scorers (all r's between 0.998 and 0.999) and between Teachers and Experts (r = 0.999 and 0.988, respectively).

These results demonstrate that text-independent machine scoring of Words Correct for children's classroom reading predicts human scores extremely well (almost always within a word or two).

Acknowledgments

The authors acknowledge useful feedback from the anonymous reviewers that improved this paper.

References

Jennifer Balogh, Jared Bernstein, Jian Cheng & Brent Townshend. 2005. Ordinate Scoring of FAN in NAAL Phase III: Accuracy Analysis. Ordinate Corporation: Menlo Park, California.

Jennifer Balogh, Jared Bernstein, Jian Cheng, Alistair Van Moere, Brent Townshend, Masanori Suzuki. In press. Validation of automated scoring of oral reading. Educational and Psychological Measurement.

Jared Bernstein, Michael Cohen, Hy Murveit, Dmitry Rtischev, Dmitry, and Mitch Weintraub. 1990. Automatic evaluation and training in English pronunciation. In: Proc. ICSLP-90: 1990 Internat. Conf. on Spoken Language Processing, Kobe, Japan, pp. 1185–1188.

Matthew Black, Joseph Tepperman, Sungbok Lee, Patti Price, and Shrikanth Narayanan. 2006. Automatic detection and classification of disfluent reading miscues in young children's speech for the purpose of assessment. Proc. In INTERSPEECH/ICSLP, Antwerp, Belgium.

Jian Cheng & Jianqiang Shen. 2010. Towards Accurate Recognition for Children's Oral Reading Fluency. IEEE-SLT 2010, 91-96.

Jian Cheng & Brent Townshend. 2009. A rule-based language model for reading recognition. SLaTE 2009.

Lindy Crawford, Gerald Tindal, & Steve Stieber. 2001. Using Oral Reading Rate to Predict Student Performance on Statewide Achievement Tests. Educational Assessment, 7(4), 303-323.

Maxine Eskanazi. 2009. An overview of spoken language technology for education. Speech Communication, 51, 832-844.

David LaBerge & S. Jay Samuels. 1974. Toward a theory of automatic information processing in reading. Cognitive Psychology, 6(2), 293-323.

Mitchell P. Marcus, Beatrice Santorini, & Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics 19(2):313-330.

Jack Mostow, Steven F. Roth, Alexander G. Hauptmann, & Matthew Kane. 1994. A prototype reading coach that listens. In Proc. of AAAI-94, 785–792.

National Institute of Child Health and Human Development, "Report of the national reading panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction," Tech. Rep. NIH Publication No. 00-4769, U.S. Government Printing Office, 2000.

Timothy V. Rasinski & . James V. Hoffman. 2003. Theory and research into practice: Oral reading in the school literacy curriculum. Reading Research Quarterly, 38, 510-522.

Joseph Tepperman, Matthew Black, Patti Price, Sungbok Lee, Abe Kazemzadeh, Matteo Gerosa, Margaret Heritage, Abeer Always, and Shrikanth Narayanan. 2007. A Bayesian network classifier for word-level reading assessment. Proceedings of ICSLP, Antwerp, Belgium.

Steve Young, D. Ollason, V. Valtchev, & Phil Woodland. 2002. The HTK Book (for HTK Version 3.2). Cambridge University Engineering Department.

Klaus Zechner, John, Sabatini, & Lei Chen. 2009. Automatic scoring of children's read-aloud text passages and word lists. Proceedings of the NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications. Boulder, Colorado.