Human and Automated Assessment of Oral Reading Fluency

Article *in* Journal of Educational Psychology · November 2013

DOI: 10.1037/a0031479

6 authors, including:

Daniel Bolaños
Corporacion universitaria Unicomfacauca

Ronald A. Cole
Boulder Learning Inc.

Wayne Ward
University of Colorado Boulder

Gerald Tindal
University of Oregon

Human and Automated Assessment of Oral Reading Fluency

Daniel Bolaños, Ron A. Cole, Wayne H. Ward

Boulder Language Technologies, Boulder CO

Gerald A. Tindal

University of Oregon, Eugene, OR

Jan Hasbrouck

Gibson Hasbrouck & Associates Wellesley MA

Paula J. Schwanenflugel

The University of Georgia, Athens, GA

Author Note

Manuscript submitted to the Journal of Educational Psychology, Special Issue on "Advanced Learning Technologies"

Abstract

This article describes a comprehensive approach to fully automated assessment of children's Oral Reading Fluency (ORF), one of the most informative and frequently administered measures of children's reading ability. Speech recognition and machine learning techniques are described that model the three components of oral reading fluency: word accuracy, reading rate and expressiveness. These techniques are integrated into a computer program that produces estimates of these components during a child's one-minute reading of a grade-level text. The ability of the program to produce accurate assessments was evaluated on a corpus of 783 one-minute recordings of 313 students reading grade-leveled passages without assistance. Established standardized metrics of accuracy and rate (Words Correct Per Minute (WCPM)) and expressiveness (National Assessment of Educational Progress expressiveness scale) were used to compare ORF estimates produced by expert human scorers and automatically generated ratings. Experimental results showed that the proposed techniques produced WCPM scores that were within 3 to 4 words of human scorers across students in different grade levels and schools. The results also showed that computer-generated ratings of expressive reading agreed with human raters better than the human raters agreed with each other. The results of the study indicate that computer generated ORF assessments produce an accurate multidimensional estimate of children's oral reading ability that approaches agreement among human scorers. The implications of these results for future research and near term benefits to teachers and students are discussed.

Keywords: oral reading fluency, automated reading assessment, expressive reading, automatic speech recognition.

Human and Automated Assessment of Oral Reading Fluency

Reading assessments provide school districts and teachers with critical and timely information for identifying students who need immediate help, for making decisions about reading instruction, for monitoring individual student's progress in response to instructional interventions, for comparing different approaches to reading instruction, and for reporting annual outcomes in classrooms, schools, school districts and states. One of the most common tests administered to primary school students is *oral reading fluency*. Over 25 years of scientifically-based reading research has established that fluency is a critical component of reading and that effective reading programs should include instruction in fluency (Kuhn and Stahl, 2000; Fuchs et al., 2001; National Reading Panel, 2000). While oral reading fluency does not measure comprehension directly, there is substantial evidence that estimates of ORF predict future reading performance and correlate strongly with comprehension (Fuchs et al., 2001; Shinn, 1998). According to Wayman et al. (2007), ORF is valid indicator of comprehension in early grades, though less so beyond grade 4. Because oral reading fluency can be measured rather quickly (typically in 5 to 10 minutes) with good validity and reliability, it is widely used to screen individuals for reading problems and to measure reading progress over time.

In this article we present a comprehensive approach to assess ORF automatically through the use of speech recognition and machine learning techniques. The approach is comprehensive because all three measures of ORF, accuracy, rate (combined into a WCPM score) and expressiveness can be measured automatically and in real time, whereas expressiveness is rarely scored in real world educational contexts. The ultimate goal of automatic assessment of ORF is to provide an effective and low cost alternative to human-administered assessments. in order to reduce the millions of hours of time teachers spend each year assessing their students' reading abilities, which is mandated by federal law in the U.S. In addition, computer-based assessments of ORF could generate detailed records of individual student's performance, including the digital recordings of each reading session, that could be reviewed by teachers, parents and students, and analyzed automatically for detailed information about the student's reading problems. Automatic administration of ORF will also enable collection of massive amounts of speech data that can be used to analyze and understand children's development of reading skills; these data can also be used to improve the performance of the speech recognition technologies.

We used a speech recognition system specifically designed to process children's read speech to produce a word-level hypothesis of what the student read from a grade-level text during one minute. From this hypothesis and the text passage, a Words Correct Per Minute (WCPM) score was computed reflecting the student's reading accuracy and rate. In order to assess prosodic reading, we developed a series of lexical and prosodic features that were extracted from the student's speech. These included analysis of the text syntax and its correlation with filled-pauses and silence regions, syllable and word duration, pitch, and word co-occurrences, among other features described below. Machine learning classifiers were trained on these features, resulting in statistical models that were able to discriminate between different degrees of prosodic reading using the NAEP ORF Scale (Daane et al., 2005). A hierarchical classification scheme was utilized in order to assign one-minute reading sessions to levels in the NAEP scale.

The accuracy of these assessment methods was evaluated on approximately 13 hours of speech collected from the 313 first through fourth grade students who read grade level text passages. WCPM scores as well as NAEP assessments generated by the system, FLORA (FLuent Oral Reading Assessment) were compared to those produced by at least two independent human judges.

The remainder of the article is organized as follows: the next section provides the scientific rationale for assessing oral reading fluency. We then describe the corpus of children's read speech that was collected for this study. We then describe the system and features used to assess WCPM (accuracy and rate) and expressive reading using lexical and prosodic features extracted from the speech. The last section presents the discussion and conclusions.

Scientific Rationale for FLORA

Oral Reading Fluency.

Oral Reading Fluency (ORF) is typically defined as a student's ability to read words in grade level texts accurately and effortlessly, at a natural speech rate and with appropriate prosodic expression. A synthesis of scientifically-based reading research by the National Reading Panel [18] concluded that "Reading fluency is one of several critical factors necessary for reading comprehension, but it is often neglected in the classroom. If children read out loud with speed, accuracy and proper expression, they are more likely to comprehend and remember the material than if they read with difficulty and in an inefficient way."

Accuracy and Automaticity. Accurate reading speed is both a strong discriminator of reading ability (e.g., Perfetti, 1985; Jenkins, et al., 2003), and a strong predictor of later reading proficiency (Lesgold & Resnick, 1982; Scarborough, 1998; see review by Compton and Carlisle, 1994.) As Jenkins et al. (2003) put it: "Together with listening comprehension, word-reading skill accounts for nearly all of the reliable variance in reading ability, and individual differences in word recognition explain significant variance in reading ability, even after controlling for reading comprehension (Curtis, 1980; Hoover & Gough, 1990)."

Oral reading fluency depends upon the ability to recognize words in a text *quickly and automatically*. As defined by Fuchs et al. (2001), automaticity is "the oral translation of text with speed and accuracy." Automaticity theory (LaBerge & Samuels, 1974; Samuels, 1985; Wolf, 1999) and related verbal-efficiency accounts of reading (Perfetti, 1985) hold that students who have learned to decode printed words automatically are able to devote more attention (cognitive resources) to comprehending what they are reading. Readers who have not achieved automaticity during word recognition must devote significant attention to recognizing words (at the expense of devoting this attention to making sense of the text), resulting in slower reading times and weaker comprehension. Support for automaticity and the verbal-efficiency theories of reading is provided by the strong association between the speed of reading words, either in word lists or in context, and measures of reading comprehension.

Expressiveness. While readers who have achieved fluency can read texts rapidly and accurately, they may not read expressively, i.e., they may not pause between sentences, at major phrase boundaries within sentences, or produce appropriate prosody when reading out loud. Expressive reading is the third critical component of reading fluency, typically defined as reading a text with the appropriate expression, intonation and phrasing in order to preserve meaning (Miller & Schwanenflugel, 2008).

Connection between ORF and Comprehension. For over 25 years, researchers have documented the association between reading fluency and comprehension. Reviews of the research on ORF have demonstrated consistently moderate to strong correlations between ORF and comprehension (Marston, 1989; Shinn, 1998). Research results have demonstrated high concurrent validity between ORF and measures of word recognition and reading comprehension (Hosp & Fuchs, 2005; Jenkins et al., 2003), and between ORF and nationally normed standardized tests of reading comprehension (Roehrig et al., 2008; Schilling et al., 2007; Schwanenflugel et al., 2006). Measures of ORF in early grades have also been found to predict comprehension in later grades. (Kim et al., 2010). Thus, the relation between ORF and reading comprehension has been well established by previous research, particularly for students in elementary school (Roberts et al, 2005; Roehrig et al., 2008; Kim et al., 2010).

Previous Work using Automatic Speech Recognition to Assess and Improve ORF.

Automatic Assessment of Reading Accuracy and Rate. Over two decades of research has investigated the use of automatic speech recognition (ASR) to assess and improve reading. Seminal research conducted by Jack Mostow and his colleagues in Project Listen at Carnegie Mellon University has demonstrated the effectiveness of ASR for improving reading fluency and comprehension for both native and nonnative speakers of English (Mostow et al. 2003; Reeder et al. 2007). Mostow et al. (2003) used an ASR system to measure a student's interword latency, defined as the elapsed time between certain words read aloud by the student that were scored as correctly read by the ASR system. Their model of interword latency produced a correlation of over 0.7 with independent WCPM measures of oral reading fluency using grade level passages.

In the context of Project Tball (Technology Based Assessment of Language and Literacy) at UCLA and USC, Black et al. (2008) investigated oral reading of 55 isolated words produced by kindergarten, 1st and 2nd grade children with the aim of detecting reading miscues automatically, such as sounding-out, hesitations, whispering, elongated onsets, and question intonations. Black et al. developed an ASR system that used specialized grammars to model word-level disfluencies using the subword modeling approach developed by Hagen and Pellom (2005). Scores produced by the recognition system correlated highly (.91) with fluency judgments provided by human listeners.

A series of studies by Bryan Pellom and Andreas Hagen and their collaborators (Hagen et al. 2007) investigated ways to optimize an ASR system for children's read speech. The research resulted in a reduction in the word error rate (WER) from 17.4% to 7.6%. Hagen et al. (2007) developed a version of the ASR system that used subword modeling rather than whole word scoring to detect reading errors. In the study several subword lexical units and approaches were evaluated for detection of reading disfluencies and modest gains were reported. Bolaños (2008) reported that additional detection gains were achieved by using syllable graphs to represent hypotheses from the ASR system.

Automatic Assessment of Expressive Oral Reading. Although the National Reading Panel (2000) and research community define oral reading fluency in terms of word recognition accuracy, reading rate and how expressively the student reads (see Kuhn et al, 2010; for a discussion of this topic), expressiveness is rarely measured in assessments of ORF. Only recently has the expressiveness aspect of the reading fluency construct found its way into automated assessments of fluency. Duong et al. (2011) investigated two alternative methods of measuring prosody during children's oral reading. The first method, which was text-dependent, consisted of generating a prosodic template model for each sentence in the text. The template was based on word-level features like pitch, intensity, latency and duration extracted from fluent adult narrations. The second method investigated adult narrations to train a general duration model that could be used to generate expected prosodic contours of sentences for any text, so an adult reader was no longer required to generate sentence templates for each new text. Both methods were evaluated for their ability to predict student's scores on fluency and comprehension tests, and each produced promising results, with the second, automated method for generating prosodic sentence templates outperforming adult narrations of each individual text. However, none of these methods could satisfactorily classify sentences using the NAEP expressiveness rubric which was probably due to the low human inter-rater reliability reported in this study.

Development of the FLORA System

Development of a Corpus for Assessing Oral Reading Fluency

Data Collection Setting. Data were collected from 313 first through fourth grade students in four elementary schools (9 classrooms) in the Boulder Valley School District (BVSD) in Colorado. Data were collected from students in their classrooms at their schools. School 1 scored proficient or above on the state reading assessment.

> School 2 had 51.7% students with free or reduced lunch (similar to School 1), but 79% of third grade students tested as proficient or above on the state literacy test. School 2 was a bilingual school with nearly 100% English learners (ELs) who spoke Spanish as their first language. School 3 had 18.4% of students with free or reduced lunch, 85% of students were proficient or above in the state literacy test. School 3 also had relatively few ELs.

Text Passages. Twenty text passages were available for reading at each grade level. The standardized text passages were downloaded from a website (Good et al., 2007) and are freely available for noncommercial use. The passages were designed to assess ORF and are about the same level of difficulty at each grade level. ORF norms have been collected for these text passages for tens of thousands of students at each grade level in fall, winter and spring semesters, so that students can be assigned to percentiles based on national WCMP scores (Hasbrouck and Tindal, 2006).

Data Collection Protocol. The data were collected using the FLORA system (Bolaños et al., 2011), which was configured to enroll each student, randomly select one passage from the set of 20 standardized passages for the student's grade level, and present the passage to the student for reading out loud. Because testing was conducted in May, near the end of the school year, classroom teachers had recently assessed their student's oral reading performance (using text passages different from those used in our study). About 20% of the time, teachers requested that specific students be presented with text passages either one or two levels below or one or two levels above the student's grade level. Thus, about 80% of students in each grade read passages at their grade level, while 20% of students read passages above or below their grade level, based on their teachers' recommendations. Depending upon the number of students that needed to be tested on a given day, each student was presented with two or three text passages to read aloud.

During the testing procedure, the student was seated before a laptop, and wore a set of headphones with an attached noise-cancelling microphone. The experimenter observed or helped the student enroll in the session, which involved entering the student's gender, age and grade level. FLORA then presented a text passage, started the one minute recording at the instant the passage was displayed, recorded the student's speech and relayed the speech to a server.

Corpus summary. The corpus comprised 783 recordings from 313 first through fourth grade students for a total of approximately 13 hours of speech data. Each recording was scored manually be two human judges. Words were scored as reading errors if the word was skipped over, or the judge decided that the word was misread. Insertions of words (intrusions) were not scored as reading errors, as insertions were not counted as errors in the national norms collected by Hasbrouck and Tindal (REF).

Automatic Generation of WCPM Scores

The number of words that a student read correctly during one minute was computed automatically by ReadToMe, the reading tracker built on top of our ASR system. ReadToMe, which resides on a server, receives the audio input in real-time from the computer in the student's classroom, computers a WCPM score for the recording. The computation of the WCPM score is done as follows. (1) ReadToMe uses an ASR system developed by Daniel Bolaños (Bolaños et al. 2011, Ward et al. 2011) to produce a word-level hypothesis representing what the student read. (2) ReadToMe aligns the hypothesis to the reference text (the story read) and tags each of the words in the reference as correctly or incorrectly read or skipped over. (3) Finally, ReadToMe counts the number of words scored as correctly read during the one minute reading; this number is the WCPM score.

Automatic Assessment of Expressive Reading

In order to assess expressive reading automatically, we proposed a set of lexical and prosodic features that can be used to train a machine learning system to classify how expressively students read text passages aloud using the 4-point NAEP scale. The proposed features were designed to measure the speech behaviors associated with each of the four levels of fluency described in the NAEP rubric, and were informed by research on acoustic phonetic, lexical and prosodic correlates of fluent and expressive reading described by other research (Kuhn et al. 2010). The proposed features were extracted from multiple sources including the recognition hypothesis, a pitch-extractor and a syllabification tool. Features included the WCPM score itself, the speaking rate, sentence reading rate, number of word-repetitions, location of the pitch accent, word and syllable durations, filled and unfilled pauses and their correlation to punctuation marks in the story read. A detailed description, motivation and analysis of all the features proposed and used for the study can be found in (Bolaños et al., 2012a).

Classification method: In order to classify the 783 one-minute recordings using the features proposed, we used a powerful classification technique called Support Vector Machine classifiers (SVMs) (Vapnick, 1995). We experimented with difference classification strategies and found a strategy based on a Decision Directed Acyclic Graph (DAG) (Platt et al., 2000) to be the most successful. The DAG approach makes sense conceptually because it maps directly to the the NAEP scale; i.e., it distinguishes disfluent from fluent speech (levels {1, 2} and {3, 4} respectively in the NAEP scale) and then makes finer distinctions ({1} vs {2} and {3} vs {4}). To implement the DAG strategy we trained three classifiers. The first classifier was trained on samples from all classes and separated samples from classes {1, 2} and {3, 4}. This classifier was placed at the root of the tree while two other classifiers, trained on samples from classes {1, 2} and {3, 4}, respectively, were placed on the leaves of the tree to make the finer grain decisions. A detailed description of the classification scheme can be found in Bolaños et al. (2012a).

Speech Recognition System

A total of 106 hours or read speech from three different children's speech corpora were used to train the recognition system. The recognizer was not trained on the corpus of read speech, described above, that was used to evaluate FLORA. We note that the system is *text independent*; that is, for new text passages the system automatically generates the expected pronunciation(s) of each word in a text passage from a pronunciation dictionary.

The speech recognition system combines two main sources of information—the scores produced by the match of the system's acoustic models to the student's speech to score each word in terms of the expected phoneme sequences extracted from a pronunciation dictionary, and the probabilities of word co-occurrences within the text (the statistical language models). These two sources of information are combined to produce the most likely hypothesis string given the speech input. Additionally, phone-level alignments from each of the one-minute recordings were generated for feature extraction purposes. Two complementary speaker adaptation techniques were utilized in order to tailor the speaker-independent acoustic models to the speech characteristics and vocal tract length of each speaker.

Comparison between Automated and Human Assessments of ORF

Human Scoring of Recorded Sessions

In order to evaluate the ability of FLORA to produce reliable WCPM scores, each of the 783 one-minute recordings collected as part of the evaluation corpus was scored independently by two former elementary school teachers. Each teacher had more than a decade of experience administering reading assessments to elementary school children. The scorers were able to listen to, review and modify their judgments within each recording until they were satisfied with their WCPM score. Thus they were allowed to listen to the recording more than once.

Additionally, each of the 783 recordings was scored from 1 to 4 using the NAEP ORF Scale by at least two independent scorers, which were former elementary school teachers with experience assessing reading proficiency. A set of 70 stories of the total 783 stories were scored by the five available teachers while the other recordings were scored by just two of them, which were randomly assigned to each scorer. A training session was scheduled before the scoring process to review the NAEP scoring instructions and unify criteria. The judges first listened to passages rated by two experienced researchers whose area of expertise is expressive reading (Paula Schwanenflugel and Melanie Kuhn). The teachers who scored the stories then rated these passages, and compared their ratings to the experts. The teachers then rated several additional passages and discussed their connection to the definition of each of the NAEP levels. This process stopped once their level of agreement approximated the agreement exhibited by the two experts.

For the actual scoring of the evaluation corpus scorers listened to each 60 second story in 20 second intervals, and provided a 1 to 4 rating for each interval. The NAEP ORF Scale (Daane et al., 2005) comprises 4 levels from less to more fluent. Level 1 is characterized by word-by-word reading, level 2 by reading using two-word phrases with some three- or four-word groupings, level 3 is characterized by a majority of three- or four-word phrase groups while preserving the the syntax of the author, finally, readers at level 4 read primarily in larger, meaningful phrase groups with expressive interpretation. Finally scorers attached a global NAEP score to the recording based on the NAEP scores assigned to each 20 second segment, which were combined using their best judgment rather than using a deterministic method like the mean or mode. A training session was held before the teachers independently scored the recordings.

Assessment of Reading Accuracy and Automaticity

Table 1 shows the mean and standard deviation (between parentheses) for Accuracy, WPM (words per minute) and WCPM scores for the human scorers and FLORA. Statistics are shown per reading level for students in the four schools. As noted above, although the evaluation data was collected from students from grades 1 to 4, about 20% of the time, teachers requested that specific students be presented with text passages either one or two levels below or above the student's grade level, resulted in reading levels for text passages from grades 1 to 6. In Table 1 accuracy is expressed in percentages and WPM, which measures fluency from the perspective of speed ignoring accuracy, is based on the average across the two human scorers for each recording. It can be seen that accuracy (percentage of words read correctly) is higher for higher grade-levels, from 70.3% for first grade to 92.6% and 90.5% for 5th and 6th grade-levels respectively. WPM are displayed in column 5 for each grade level; as expected they are highly correlated with WCPM measured by human scorers (column 5), however WCPM computed by FLORA (column 7) are much closer to human WCPM scores (column 6) than WPM. A major result can be observed by comparing the WCPM scores from the human scorers and FLORA, which present a very similar distribution (mean and standard deviation). In addition, we observed a very similar distribution of WCPM scores from humans and FLORA within each of the nine classrooms in which we conducted the study, even for classrooms in schools in which the majority of students spoken Spanish as their first language and were officially designated as English learners. Column 8 shows the expected number of WCPM for each grade level according to Hasbrouck and Tindal (2006) reading norms. It can be seen in the table that students were assigned by teachers to reading levels at which they read around the 50th percentile. We believe that there is no credible evidence to link higher WCPM scores to improved comprehension but there is substantial support for the need for readers to have an accuracy and rate (WCPM score) in the range of the 50th percentile to support both comprehension and motivation.

Another pattern of results is revealed by examining the numbers in column 9, which shows the mean difference in WCPM scores for the two human scores for the recordings in each classroom, and the numbers in column 10, which shows the mean difference between the averaged human scores and FLORA for each classroom. Note that differences in WCPM scores are expressed in absolute value. Viewing the numbers in column 9 reveals the remarkable agreement between the two human scorers (1.2 WCPM difference across all schools) and the low variance. Across all recordings, the mean difference between FLORA and the averaged human scores was 3.6 words, while the mean difference between human scores was 1.2 words.

Figure 1a displays a scatter plot of the WCPM scores from the two human scorers for all recordings, while Figure 1b displays a scatter plot of the WCPM scores from FLORA with respect to the average human scores for all recordings. If agreement were perfect, all points would lie on the diagonal. These figures show the strong agreement between WCPM scores for human scorers on each recording, and the very good agreement between FLORA and the human scores, with relatively few outliers.

We were interested in determining if FLORA might be a useful tool for providing a WCPM score that could be used as one valuable indicator, along with other measures, to identify students who at-risk for failing to learn to read. One way to do this is to compare human and FLORA WCPM scores to national reading norms developed by Hasbrouck and Tindal (2006).

The inter-rater agreement in the task of mapping recorded stories to percentiles was 0.97 for the human scorers and 0.89 between FLORA and each of the human scorers. The inter-rater agreement in the task of mapping recorded stories above/below the 50^{th} percentile (which is used normally as a reference to identify at risk students) was 0.98 for the human scorers and 0.92 between FLORA and each of the human scorers. Agreement was computed using the Weighted Kappa coefficient (κ) (Cohen, 1968) which is suitable for ordinal categories.

As can be seen the inter-human agreement and the FLORA to human agreement is very close, which means that FLORA performs well at identifying students that might require additional reading assessments and instruction.

Assessment of Expressive Reading

In this section we show results on assessing expressive oral reading using FLORA. First we briefly analyze the classification accuracy for the lexical and prosodic features proposed in relation to human assessments. We then analyze agreement and correlation between human scores and the proposed automatic scoring system using the NAEP scale.

Classification Accuracy.

In order to derive the most effective combination of features to assess expressive reading we measured the classification accuracy (percentage of recordings that FLORA assigned the same label than the human labelers) of FLORA on the corpus described above. Each recording was labeled by FLORA according to the NAEP scale and labels were compared to those from all the available human labelers. We note that there exists an upper bound to the classification accuracy that can be attained by the classifier. The reason is that whenever the human raters score the same recording differently there is an unrecoverable classification error.

Results showed that both lexical and prosodic features contributed similarly to the classification accuracy for the NAEP-2 task (89.27% and 89.02% respectively). This can be initially considered an unexpected result since lexical aspects like the number of words read correctly are expected to dominate the discrimination between fluent and non-fluent readers. However it is important to note that some of the prosodic features defined in this study are very correlated to the lexical features. For example, it is obvious that the number of words correctly read in a one-minute reading session should correlate to the average duration of a silence region or the number of filled pauses made.

For both the NAEP-2 and NAEP-4 tasks, lexical and prosodic features provided complementary information that led to an improved classification accuracy when combined. For the NAEP-4 tasks, lexical features seem to have a dominant role (73.24% and 69.73% respectively). We attribute this to the WCPM score, which is taken as a lexical feature; this score by itself provides a 71.78% accuracy for the NAEP-4 task. As expected the automatically computed WCPM, which comprises two of the three reading fluency cornerstones (accuracy and rate) plays a fundamental role. In particular the combination resulted on accuracies of 90.72% and 75.87% for the NAEP-2 and NAEP-4 tasks. Finally note that the distribution of recordings across the NAEP levels according to humans and machine was very similar.

Inter-rater Agreement and Correlation.

In this section we present inter-rater agreement and correlation results for the best system from the previous section (multi-label training using all the features). Table 2 shows the inter-rater agreement for the tasks of classifying recordings into the broad NAEP categories (fluent vs non fluent), referred as NAEP-2, and the 4 NAEP categories, referred as NAEP-4. For the

NAEP-2 task the inter-rater agreement is measured using the Cohen's Kappa coefficient (κ) (Cohen, 1960); where p(a) is the probability of observed agreement while p(e) is the probability of chance agreement.

For the NAEP-4 task we measured the inter-rater agreement using the Weighted Kappa coefficient (κ) (Cohen, 1968) which is more suitable for ordinal categories given that it weights disagreements differently depending on the distance between the categories (we used linear weightings). As a complementary metric for this task we have computed the Spearman's rank correlation coefficient (Spearman, 1904). In a number of classification problems, like emotion classification, the data is annotated by a group of human raters who may exhibit consistent disagreements on similar classes or similar attributes. In such classification tasks it is inappropriate to assume that there is only one correct label since different individuals may consistently provide different annotations (Steidl et al., 2005). While the NAEP scale is based on clear descriptions of reading behaviors at each of four levels, children's reading behaviors can vary across these descriptions while reading, and individuals scoring the stories may differ consistently in how they interpret and weight children's oral reading behaviors. For this reason, we believe that examining correlations between human raters and between human raters and the machine classifiers is a meaningful and useful metric for this task.

Each row in the table shows the agreement and correlation coefficients of each rater respect to the other raters (excluding FLORA in the case of the human raters), note that not all the scorers scored the same number of recordings. In order to interpret the computed Kappa values, we have used as a reference the interpretation of the Kappa Coefficient provided in Landis and Koch (1977), which attributes *good* agreement to Kappa values within the interval

[0.61−0.80] and *very good* agreement to higher Kappa values [0.81−1.00]. According to this interpretation Table 6 reveals that: a) there is *good* inter-human agreement for both the NAEP-2 and NAEP-4 tasks, b) there is *good* FLORA-to-human agreement for the NAEP-4 task, and c) there is *very good* FLORA-to-human agreement for the NAEP-2 task. It can be observed that the Kappa agreement between FLORA and the humans is higher than the agreement between each human scorer and the rest of the human scorers. This is true for both the NAEP-2 and NAEP-4 tasks. This difference in agreement is statistically significant, which shows the ability of the proposed features and classification scheme to provide a useful method to automatically assess expressive oral reading using the NAEP scale.

In terms of the Spearman's rank correlation coefficient (ρ) we obtained relatively strong inter-human correlation ([.80-.81]) and an even stronger machine-to-human correlation (.86) in the NAEP-4 task. This indicates that NAEP-scores from every pair of scorers are closely related, which is consistent with the weighted Kappa values obtained.

In table 3 we display cross-tabs of agreement/disagreement between humans and between FLORA and humans (in percentages). In both cases most of the data lies in the main diagonal and we believe that there are no obvious biases between humans and FLORA.

Connection between Reading Accuracy, Reading Rate and Expressive Reading.

We conducted a set of analysis to gain insights into the relationship between the two main measures of oral reading fluency—WCPM and expressiveness. These analyses are displayed in Figure 2a and 2b. In each of these figures, we sorted students according to their WCPM percentile using the Hasbrouck and Tindal (2006) norms. Thus, the bar at leftmost of each figure represents students with WCPM scores below the 10th percentile, whereas the rightmost bar shows students in the 90th percentile. Figure 2a displays percentile assignments based on average human scorers rating, and Figure 2b displays percentile assignments based on FLORA WCPM estimates. The colors within each bar indicate the percentage of students at each NAEP score; in Figure 2a these numbers are based on the NAEP scores assigned by the human scorers, and in Figure 2b these numbers were assigned by FLORA.

It is clear from this figure that recordings in the highest percentiles (highest reading accuracy and rate) correspond to more expressive readers (higher levels in the NAEP scale). For example, all of the recordings for students in the 90th percentile based on WCPM were assigned to levels 3 and 4 in the NAEP scale. Moreover, about 97.0% of the recordings below the 10th percentile were assigned to levels 1 and 2 in the NAEP scale. Figures 2a and 2b reveal several interesting patterns: a significant percentage of recordings placed below the 50th percentile (which might be used to identify students in need for fluency support) were placed in the higher levels of the NAEP scale according to our expert human annotators (3.08%, 24.02% and 45.19% for recordings below the 10th percentile, in the 10th percentile and in the 25th percentile respectively). This means that there are a number of speakers who, despite reading below the expected rate according to the percentiles published by Hasbrouck and Tindal (2006), read with appropriate/good expression and would be considered fluent readers according to the NAEP scale. Another interesting observation is that a significant percentage of recordings placed above the 50th percentile were assigned to the lower levels in the NAEP scale by our expert human annotators. Those recordings likely correspond to speakers that are reading for speed rather than for comprehension in order to get as many words read as possible within the one minute session. In particular 24.88% of the recordings in the 50th percentile were assigned to levels 1 and 2 in the

NAEP scale (non fluent) while 13.92% of the recordings in the 75th percentile were assigned to those levels. We note that the instructions provided to students before recording stories emphasized the importance of reading the text naturally, rather than as fast as they could; these percentage might have been higher if we had not emphasized reading naturally in the instructions. These observations suggest that it measuring expressiveness as well as WCPM is likely to be both informative and beneficial to understanding individual student's oral reading abilities. Finally, we note that Figure 5, which is analogous to Figure 4 but was built using FLORA scores, presents very similar information.

Discussion and Conclusions

We investigated the automatic assessment of Oral Reading Fluency in children's speech according to two standard rubrics: WCPM (to measure accuracy and rate) and the NAEP Expressiveness scale. Compared to human scoring of WCPM and expressiveness on 783 one-minute recordings of children reading grade-level text passage, results show that automatically generated WCPM scores differ by an average of 3.5 words with respect to the human-average score for each recorded story, while humans differ by an average of 1.5 words for each story.

For expressiveness, FLORA had an accuracy of 90.93% classifying recordings according to the binary NAEP scale ("fluent" versus "non-fluent") and 76.05% on the more difficult 4-point NAEP scale. According to the classification of Kappa strength proposed by Landis and Koch (1977), the Kappa agreement for both NAEP-2 and NAEP-4 tasks between each human scorer and the rest of the human scorers was *good*, while the Kappa agreement between the machine and the human scorers was *good* and *very good* respectively. the. In addition, the Kappa agreement between FLORA and each human scorer was always significantly higher than the

Kappa agreement between the human scorers. In terms of the Spearman's rank correlation coefficient (ρ), correlation between the machine and each human scorer was always significantly higher than the correlation between human scorers.

The results of the research reveal that speech recognition and machine learning systems can produce accurate assessments of WCPM and expressiveness that approach (WCPM) or exceed human performance. Without question, the results of the WCPM scores reported above can be improved substantially in the near future using known ASR solutions, such as collecting more training data to model children's speech patterns. For example in Vergyri1 et al. (2010) it is reported that accent-dependent acoustic modeling (which implies training/adapting on data from the target accent) produces a significant increase in recognition performance compared to accent-independent modeling. In a recent study that we conducted on 191 native Spanish children learning to read English text in Spanish schools (Bolaños et al., 2012b), we determined experimentally that statistical models trained on speech from the target population were significantly more accurate than models trained on native English children. Results from that study showed a mean difference in WCPM scores of 5.49 and 4.96 between FLORA and each of the human scorers, while the mean difference between the human scorers was about 5.92 words.

Perhaps the major limitation of this study is the small number of students (313) used in our research (a few hundred). To fully demonstrate the feasibility and validity of a fully automatic assessment of oral reading fluency, speech data during oral reading of leveled texts must be collected for a large and diverse population of students at different grade levels, representing students with different dialects and accents. The system must also be tested with data collected from many different classrooms or computer labs to model the acoustic environments and the realities of real word use in schools.

Towards Valid Automatic Assessment of Oral Reading Fluency

We believe there are great potential benefits of incorporating measures of expressiveness into assessments of oral reading fluency. One of the major criticisms of using WCPM to measure individual student's improvements in reading over time, (that is, in response to instruction) is that students strive to read texts as quickly as possible in order to increase their WCPM scores, which teachers often set as learning targets within a reading instruction program. When a student's ability is measured in terms of how quickly they can read the words in a text, teachers and students learn to focus on fast reading, rather than reading the text at a normal reading rate with good expression that communicates the meaning of text, and thus reflects its comprehension by the student. Fast readers have short segment durations, muted stress marking, and reduced phrase-final bracketing than slow readers, so the normal comprehension benefits children might experience by reading with good prosody may not be derived by students who are trying to read fast (Kuhn et al., 2010; Benjamin & Schwanenflugel, 2010). In sum, the emphasis on speed that can result from using WCPM as a measure of reading achievement may undermine the goal of helping students develop strategies for reading with deep understanding.

Incorporating measures of expressiveness into assessments of oral reading fluency could mitigate this problem. One can easily imagine a weighted measure of ORF that combines WCPM and expressiveness estimates, such that students receive the highest score when the words in a text are read at a natural speaking rate with prosody appropriate to the discourse structure of the text. In fact, some rating systems of reading expressiveness such as the Multidimentional Fluency Guide (Rasinski, Rikli, & Johnston, 2009) already do this.

One of the major benefits of the automated scoring of reading prosody by FLORA that neither the NAEP nor the other various teacher rating systems for evaluating reading fluency have is that these reading fluency scales have not (as yet) been grounded in research on reading prosody. We do not know whether the ratings obtained using these scales would be spectrographically valid, that is, that children rated as expressive on these scales would be the same ones who would appear expressive when their readings are viewed on a spectrogram. Because the features used in FLORA to classify expressive reading were derived directly from spectrographic measures derived from children's speech (Kuhn et al. 2010), FLORA can make this claim. Conversely, because the teacher NAEP ratings match the spectrographic distinctions made by FLORA, FLORA has also served to validate teacher impressions of reading prosody as determined by the NAEP. In sum, fully automatic assessment or ORF that combines its three components appears to be feasible with today's technologies. Additional research is needed to determine how to use these measures to provide the most useful feedback to teachers and students to assess students' reading abilities and inform instruction.

Acknowledgments

We gratefully acknowledge the help of Angel Stobaugh, Director of Literacy Education at Boulder Valley School District, and the principals and teachers who allowed us to visit their schools and classrooms. We appreciate the amazing efforts of Jennifer Borum, and the efforts of Linda Hill, Suzan Heglin and the rest of the human experts who scored the text passages. This work was supported by U.S. Department of Education award number R305B070434, National Science Foundation award number 0733323 and NIH award number R43 DC009926-01.

References

Black, M., Tepperman, J., Lee, S., & Narayanan, S. (2008). *Estimation of Children's Reading Ability by Fusion of Automatic Pronunciation Verification and Fluency Detection.* Paper presented at the Interspeech, Brisbane, Australia.

Bolaños, D., Cole, R. A., Ward, W., Borts, E., & Svirsky, E. (2011). FLORA: Fluent oral reading assessment of children's speech. *ACM Trans. Speech Lang. Process., 7*(4), 1-19.

Bolaños, D., Cole, R. A., Ward, W., Tindal, G., Schwanenflugel, P. & Kuhn, M. (2012a) Automatic Assessment of Expressive Oral Reading. (Submitted to *Speech Communication* in December 2011, the current status is "revise and resubmit").

Bolaños, D., Elhazaz, P., Ward, W. and Cole, R. (2012b). Automatic Assessment of Oral Reading Fluency for Native Spanish ELL Children . WOCCI 2012.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 37-46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 213-220.

Compton, D. L., & Carlisle, J. F. (1994). Speed of word recognition as a distinguishing characteristic of reading disabilities. *Educational Psychology Review, 6*(2), 115-140.

Curtis, M. E. (1980). Development of components of reading skill. *Journal of Educational Psychology, 72*(5), 656-669. doi: 10.1037/0022-0663.72.5.656

Daane, M.C., Campbell, J.R., Grigg, W.S., Goodman, M.J., and Oranje, A. (2005). Fourth-Grade *Students Reading Aloud: NAEP 2002 Special Study of Oral Reading* (NCES 2006-469).

U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. Washington, DC: Government Printing Office.

Duong, M., Mostow, J., & Sitaram, S. (2011). Two methods for assessing oral reading prosody. *ACM Transactions on Speech and Language Processing (Special Issue on Speech and Language Processing of Children's Speech for Child-machine Interaction Apps), 7*(4).

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*, 239-256.

Good, R. H., Kaminski, R. A., & Dill, S. (2007). Dynamic indicators of basic early literacy skills 6th Ed., Dibels oral reading fluency. Eugene, OR: University of Oregon.

Hagen, A., Pellom, B., & Cole, R. (2007). Highly accurate children's speech recognition for interactive reading tutors using subword units. *Speech Communication, 49*(12), 861-873.

Hasbrouck, J., & Tindal, G. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher, 59*, 636-644. doi: 10.1598/RT.59.7.3

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2*(2), 127-160. doi: 10.1007/bf00401799

Hosp, M. K., & Fuchs, L. S. (2005). Using CBM as an indicator of decoding, word reading, and comprehension: Do the relations change with grade? *School Psychology Rev., 34*, 9-26.

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C. L., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology, 95*, 719-729.

Kim, Y.-S., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology, 102*, 652-667.

Kuhn, M., & Stahl, S. (2000). Fluency: A review of developmental and remedial practices. Ann Arbor, MI: Center for the Improvement of Early Reading Achievement.

Kuhn, M. R., Schwanenflugel, P. J., & Meisinger, E. B. (2010). Aligning Theory and Assessment of Reading Fluency: Automaticity, Prosody, and Definitions of Fluency. *Reading Research Quarterly, 45*(2), 230-251.

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

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Lesgold, A. M., & Resnick, L. B. (1982). How reading disabilities develop: Perspectives from longitudinal study. In J. P. Das, R. Mulcahy & A. Wall (Eds.), *Theory and Research in Learning Disability.* New York: Plenum.

Marston, D. (1989). Curriculum-based measurement: What is it and why do it? In M. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18-78). New York, NY: Guilford Press.

Miller, J., & Schwanenflugel, P. J. (2008). A longitudinal study of the development of reading prosody as a dimension of oral reading fluency in early elementary school children. *Reading Research Quarterly, 43*(4), 336-354. doi: 10.1598/rrq.43.4.2

Mostow, J., Aist, G., Burkhead, P., Corbett, A., Cuneo, A., Eitelman, S., Tobin, B. (2003). Evaluation of an Automated Reading Tutor that Listens: Comparison to Human Tutoring and Classroom Instruction. *Journal of Educational Computing Research, 29*(1), 61-117.

NIH, N. I. o. C. H. a. H. D. (2000). Teaching children to read: an evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups *Report of the National Reading Panel*. Washington, DC: National Institute of Health.

Perfetti, C. (1985). *Reading ability*. Oxford, England: Oxford University Press.

Platt, J. C., Cristianini, N., & Shawe-taylor, J. (2000). Large margin DAGs for multiclass classification *Advances in Neural Information Processing Systems* (547-553): MIT Press.

Rasinski, T., Rikli, A., & Johnston, S. (2009). Reading Fluency: More Than Automaticity? More Than a Concern for the Primary Grades? *Literacy Research and Instruction, 48*(4), 350-361.

Reeder, K., Shapiro, J., & Wakefield, J. (2007). *The effectiveness of speech recognition technology in pro-moting reading proficiency and attitudes for canadian immigrant children.* Paper presented at the European Conference on Reading.

Roberts, G., Good, R., & Corcoran, S. (2005). Story retell: A fluency-based indicator of reading comprehension. *School Psychology Quarterly, 20*, 304-317.

Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Accuracy of the DIBELS oral reading fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology, 46*, 343-366.

Samuels, J. (1985). *Automaticity and Repeated Reading*. Lexington, MA: Lexington Books.

Scarborough, H. S. (1998). Early identification of children at risk for reading difficulties: Phonological awareness and some other promising predictors. In B. K. Shapiro, P. J.

Accardo & A. J. Capute (Eds.), *Specific reading disability: A view of the spectrum* (pp. 75-199). Timonium, MD: York Press.

Schilling, S. G., Carlisle, J. F., Scott, S. E., & Zeng, J. (2007). Are fluency measures accurate predictors of reading achievement? *The Elementary School Journal, 107*, 429-448.

Schwanenflugel, P. J., Hamilton, A. M., Kuhn, M. R., Wisenbaker, J. M., & Stahl, S. A. (2004). Becoming a Fluent Reader: Reading Skill and Prosodic Features in the Oral Reading of Young Readers. *Journal of Educational Psychology, 96*(1), 119-129.

Shinn, M. (1998). *Advanced applications of curriculum based measurement*. NY: Guildford Press.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 72-101.

Steidl, S., Levit, M., Batliner, A., Nöth, E. & Niemann, H. (2005) Of all things the measure is man, Automatic classification of emotions and inter-labeler consistency. *ICASSP.*

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. NY: John Wiley & Sons.

Vergyri1, D. Lamel, L. and Gauvain J.L. (2010). Automatic Speech Recognition of Multiple Accented English Data. *Proceedings of Interspeech 2010.*

Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E., Vuuren, S. V., . . . Becker, L. (2011). My Science Tutor: A conversational multimedia virtual tutor for elementary school science. *ACM Trans. Speech Lang. Process., 7*(4). doi: 10.1145/1998384.1998392

Wayman, M.M., Wallace, T., Wiley, H.I., Tichdt, R., & Espin, C.A. (2007). Literature synthesis on curriculum-based measurement in reading. The Journal of Special Education, 41(2), 85-120.

Wolf, M. (1999). What time may tell: Towards a new conceptualization of developmental dyslexia. *Annals of Dyslexia, 49*, 3-28. doi: 10.1007/s11881-999-0017-x

Table 1

Summary of Accuracy, WPM and WCPM according to Human scorers (H) and FLORA (F).

Expected WCPM (E) are also shown.

level	#stu	#rec	acc (%)	H-WPM	H-WCPM	F-WCPM	E-WCPM	H-diff	FH-diff
1	68	171	70.3 (19.7)	54.6 (25.5)	41.9 (26.4)	42.5 (25.9)	53	1.2 (1.8)	2.7 (2.7)
2	97	242	84.6 (10.1)	99.3 (31.9)	85.7 (33.1)	86.1 (31.8)	89	1.2 (2.0)	3.8 (4.4)
3	52	128	87.3 (7.6)	113.4 (28.1)	100.1 (29.6)	101.6 (28.0)	107	1.2 (1.4)	3.6 (2.8)
4	59	147	87.4 (8.1)	124.4 (26.6)	109.9 (27.3)	112.7 (27.3)	123	1.3 (1.8)	4.1 (3.1)
5	30	76	92.6 (3.6)	156.9 (26.4)	145.6 (26.6)	145.6 (24.5)	139	1.1 (1.2)	4.6 (4.5)
6	7	19	90.5 (14.1)	145.9 (46.1)	137.3 (49.6)	137.4 (49.1)	150	1.5 (2.0)	2.8 (2.6)
all	313	783	83.3 (14.0)	103.3 (42.2)	90.1 (43.1)	91.1 (42.6)		1.2 (1.8)	3.6 (3.6)

Table 2

Inter-rater agreement and correlation coefficients on the NAEP-scale

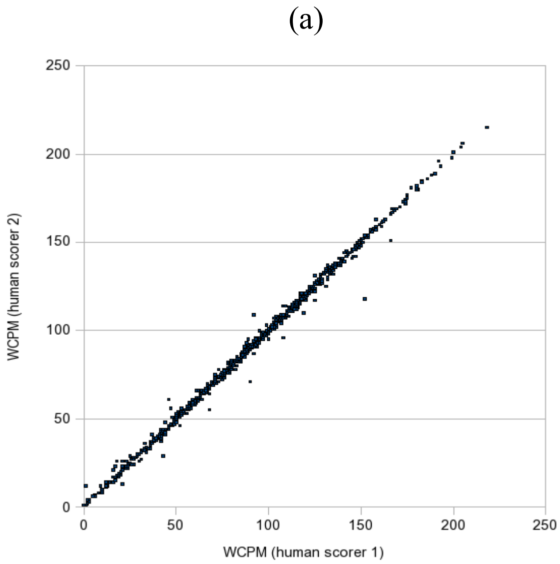
	NAEP-2				NAEP-4	
scorer	# recordings	p(a)	p(e)	κ	κ	ρ
Human 1	571	0.87	0.50	0.73	0.66	0.80
Human 2	391	0.90	0.50	0.80	0.69	0.81
Human 3	698	0.87	0.50	0.74	0.68	0.81
Human 4	799	0.86	0.50	0.71	0.69	0.81
Human 5	367	0.86	0.50	0.71	0.68	0.80
FLORA	1776	0.94	0.50	0.84	0.77	0.86

Table 3

Cross-tabs of agreement/disagreement between FLORA and human generated NAEP scores (in %).

		FLORA			
		1	2	3	4
Human	1	16.6	2.9	0.1	0
	2	3.5	21.3	3.9	0.2
	3	0	3.9	32.4	5.6
	4	0	0	3	6.6

		Human			
		1	2	3	4
Human	1	14.9	4.2	0.1	0
	2	4.4	19.6	5.7	0.2
	3	0.1	7.2	27.7	5.2
	4	0	0	4.7	6.1



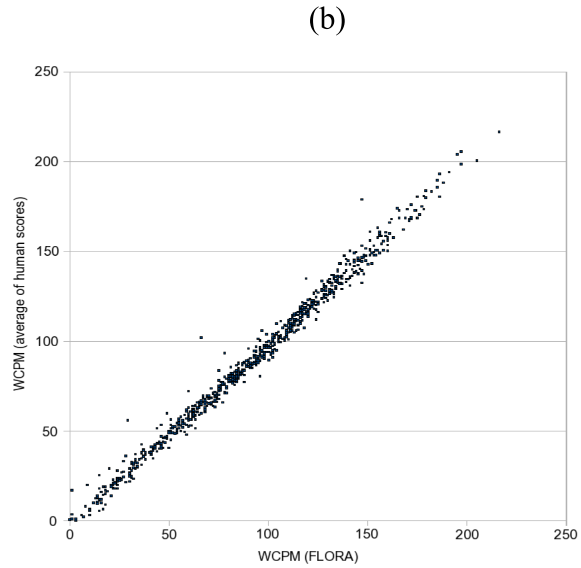


Figure 1. Correlation between WCPM scores produced by two independent human scorers (a) and between FLORA and the average of the two independent human scorers (b) for each of the one-minute recordings assessed.

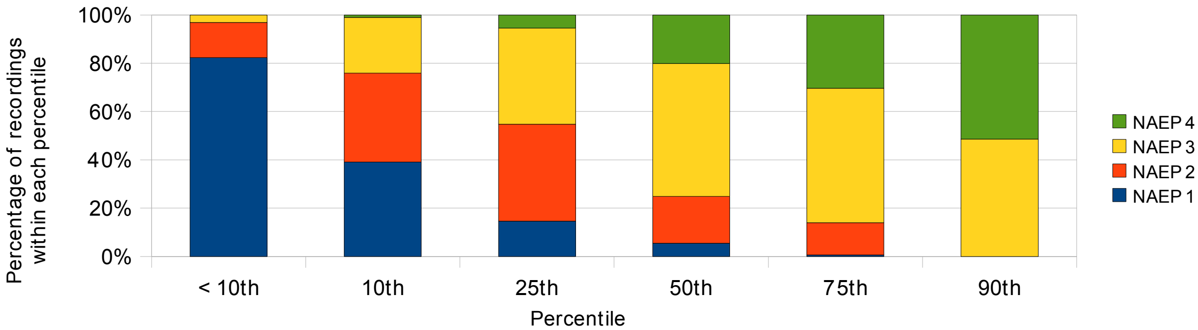


Figure 2a. Distribution of recordings across the NAEP scale for each WCPM percentile according to human scorers.

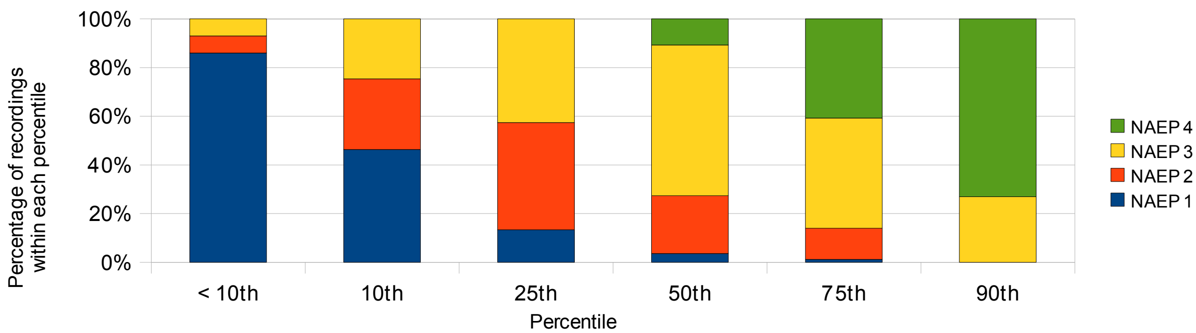


Figure 2b. Distribution of recordings across the NAEP scale for each WCPM percentile according to FLORA.