Toward a Reliable Measure of Prosody: An Investigation of Rater Consistency

Tara Haskins
Vincent Aleccia

Department of Education
Eastern Washington University
Cheney, Washington

Abstract

Reliable and valid measures of oral reading fluency are widely used in elementary language arts classes, but there are relatively few measures of prosody defined as the characteristics of pitch, stress, and phrasing students use to convey their understanding of text. Most publishers have reported limited reliability data of the few prosody-measuring instruments available. This article examines the efficacy of the WOW Prosody Rubric. This instrument uses two skills categories: Phrasing and Expression, each divided into several dimensions. Correlation coefficients between random pairs of raters selected for analysis and determined using a Spearman rho, revealed a range from low (0.2973) to moderate (0.5277) correlation. Although these coefficients are less than ideal, this study indicates that it's possible to create an instrument to assess prosody. The researchers suggest refining this instrument and then doing further field testing will increase the reliability of this measure of students' mastery of prosody.

Key Words: prosody, fluency

A fluent reader can be defined as one who reads with ample rate, with relatively few errors, and with appropriate phrasing and expression. According to the National Reading Panel (2000), all three are significant factors when examining fluency. Fluent reading is observable when readers express meaning through appropriate prosody such as pitch, stress, and phrasing (Dowhower, 1991; Rasinski, Rikli, & Johnston, 2009; Schreiber, 1991).

Reliable and valid measures of oral reading fluency (rate and accuracy) are widely used and used effectively such as Dynamic Indicators of Basic Early Literacy Skills (DIBELS) (Dynamic Measurement Group. 2009), Curriculum-Based Measures (CBM), and Gray Oral Reading Test (GORT) (Pearson Education, Inc., 2012). In addition, rate and accuracy—together labeled automaticity—and its predictive relationship to comprehension have been well-researched (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Hasbrouck & Tindal, 1992; Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003b; Shinn, Good, Knutson, Tilly, & Collins, 1992). What is not as well documented is the use of reliable measures when gauging prosody skills, and the correlation between prosody and comprehension. Some examples of current measures include the National Assessment of Educational Progress's (NAEP) Oral Reading Fluency Scale (National Center for Education Statistics [NCES], 2005) and Zutell and Rasinski's Multidimensional Fluency Scale (Zutell & Rasinski, 1991). Unfortunately, in the case of most measures of prosody, publishers reported very limited reliability data. Clark and Creswell (2010) have stated,

> There are two criteria useful for assessing whether scores from an instrument are good quality: the scores need to be reliable and the scores need to be valid…Reliable means that scores from an instrument are stable and consistent. Scores should be nearly the same when researchers administer the instrument multiple times. (p. 189)

The purpose of this project was to investigate the reliability of prosody measures as related to fluency. Fluency is defined as the rate of one's reading, accuracy of one's reading, and prosody of one's reading (that is, reading with expression including voice inflection, tone, and emphasis). Prosody in included in the definition of *fluency* (rate, accuracy, and expression) (National Reading Panel [NRP], 2000). What is not as well documented is the use of reliable measures when determining prosody. Therefore, developing reliability on a prosody measure will assist educators when measuring a student's ability to read fluently.

1. Review of the Literature

Reading fluency is defined as the ability to read at a reasonable rate with accuracy and proper expression. Prosody is an essential component of reading fluency (National Reading Panel, 2000). Rate and accuracy (together labeled *automaticity*) and its predictive relationship to comprehension has been well researched as noted above. What has not been well documented is the use of reliable prosody measures. Without a reliable measure, one cannot begin to investigate how prosody relates to reading and comprehension. Prosodic reading has been researched somewhat, but there are still significant questions researchers are asking about the topic. One recurring need is the investigation of a reliable measure of prosody. The reliability data of the few current prosody measures are irresolute.

1.1 Reading Fluency

Reading is a complex skill involving many processes. Good readers read fluently and use multiple strategies to comprehend while reading. Comprehension is generally recognized as the central goal of reading. The National Reading Panel (2000) recognized five crucial components of reading instruction for the acquisition of reading skills: (a) phonemic awareness, (b) phonics, (c) fluency, (d) vocabulary, and (e) comprehension.

Fluency is recognized as an essential component of skilled reading. Unfortunately, fluency's significance in reading instruction has not been acknowledged until recently. Allington (1983) described fluency as the "most neglected" reading skill. A study conducted by the National Assessment of Educational Progress (NAEP) found that 44% of fourth-grade students were *not* fluent with grade-level texts (Pinnell et al., 1995). In the most recent study conducted by the National Center for Education Statistics (2011), oral reading fluency was not specifically reported. The National Reading Panel started publishing a survey of topics in literacy called *What's Hot* in 1997. Leading scholars from universities, reading teachers, and administrators determined the findings of the surveys. Cassidy, Valdez, and Garrett (2010) compared the five pillars of literacy with the *What's Hot* list over time. Fluency was on the hot list in 2003 and was "very hot" until 2010. There was a growing concern that students were not achieving reading fluency. The concern grew as research agreed upon the relationship between fluency and comprehension. In 2013, participants in the *What's Hot, What's Not Literacy Survey* rated fluency as *What's Hot* and *Should Not Be Hot*, as opposed to *What's Hot, What Should Be Hot* (Cassidy & Grote-Garcia, 2013). Now that fluency has become an overlooked facet of reading, the component of prosody deserves further investigation.

Fluency is defined as the ability to read at a reasonable rate (speed) with accuracy and proper expression (prosody). The first two components, rate and accuracy, have together been labeled automaticity (Adams, 1990; Armbruster, Lehr, & Osborn, 2003). According to Rasinski (2006), "Readers must be able to decode words correctly and effortlessly (automatically) and then put them together into meaningful phrases with appropriate expression to make sense of what they read" (p. 704). Kuhn, Schwanenflugel, Meisinger, Levy and Rasinski (2010) explained that "[Fluency] is demonstrated during oral reading through ease of word recognition, appropriate pacing, phrasing, and intonation" (p. 240). Unfortunately, many educators disconnect the components of reading fluency, focusing on reading rate and accuracy, and omitting prosody. Zutell and Rasinski (1991) lamented the lack of interest in reading prosody and focused their attention on interventions to improve prosodic skills. Concomitantly, they devoted considerable effort to the development of one of the earliest efforts to measure prosody.

1.1.1 Fluency and comprehension. Fluent reading is hypothesized as a bridge to comprehension, or as a precondition or antecedent to comprehension, which involves the making of meaning from text (Adams, 1990; Allington, 1983). The most commonly used measures of comprehension are question answering, passage recall, or cloze procedure (Fuchs et al., 2001). There is ample research on fluency and comprehension (e.g., Fuchs et al., 2001; Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003a; Jenkins et al., 2003b; Hasbrouck & Tindal, 1992; Shinn et al., 1992). However, most of the research has focused on the dyad of rate and accuracy as aspects of fluency, as opposed to the triad of rate, accuracy, and prosody.

1.1.2 Oral reading fluency measures. Reliable and valid measures of oral reading fluency (rate and accuracy) are widely used and used effectively, including AIMSweb Standard Reading Assessment Passages (RAPs) (NCS Pearson, Inc., 2011), DIBELS (Dynamic Measurement Group, 2009), CBM (Deno, 1985), and GORT (Pearson Education, Inc., 2012).

Rate is commonly measured as words per minute (wpm), and accuracy is typically measured as correct words per minute (cwpm). Again, prosody has been excluded from its counterparts of rate and accuracy not only in research but also in assessment.

1.2 Reading Prosody

Many definitions of prosody include both reading with appropriate expression and phrasing (e.g., Dowhower, 1991; NCES, 2005; NRP, 2000; Pinnell et al., 1995; Zutell & Rasinski (1991). Zutell and Rasinski (1991) defined proficient fluent oral reading as "(a) the reading appears fairly effortless or automatic, (b) readers group or 'chunk' words into meaningful phrases and clauses, and (c) readers use pitch, stress, and intonation appropriately to convey the meanings and feelings they believe the author intended" (p. 212).

Prosody involves appropriate "chunking" of syntactically correct phrases (Gibson & Levin, 1975; Kuhn & Stahl, 2003; Schwanenflugel et al., 2004). There is evidence to support the contention that reading fluency acquisition is contingent on learning to chunk words into appropriate phrases, a process sometimes called *parsing* or *phrasing*. (Gibson & Levin, 1975).

In addition to phrasing, scholars have acknowledged specific characteristics of prosody. According to Dowhower (1991), "prosodic features involve variations in pitch (intonation), stress (loudness), and duration (timing)" (p. 166). Others have noted six suprasegmental speech indicators related to prosodic reading: (a) pausal intrusions, (b) length of phrases, (c) appropriateness of phrases, (d) phrase-final lengthening, (e) pitch or inflection, and (f) stress or syllabic prominence (Dowhower, 1991; Miller & Schwanenflugel, 2006; Schwanenflugel, Hamilton, Kuhn, Wisenbaker, & Stahl, 2004).

Musical aspects of prosody are also emphasized in the research. Some scholars acknowledge prosody's rhythmic or dramatic element (Allington, 1983; Erekson, 2010). This can be characterized as intonation. Erekson (2010) claimed that we need to not only acknowledge syntactic prosody's word-reading and phrase-chunking, but also emphasize prosody's interpretive qualities. He synthesized definitions of the dramatic aspect of prosody by saying that "prosody is the music of speech…Prosody is the *way* we say words and phrases beyond their phonemic and lexical qualities. Phrasing in spoken and written English is based on normative musical patterns for grouping words together" (pp. 80-81). Prosody is sometimes considered the melodic component of oral reading.

1.2.1 Research on prosody. According to some scholars, a relationship exists between prosody and comprehension, but the relationship is still unclear. Several studies have found that *expressive* readers usually read at a faster rate and have a better comprehension of what they read (e.g., Benjamin & Schwanenflugel, 2010; Clay & Imlach, 1971; Klauda & Guthrie, 2008; Miller & Schwanenflugel, 2006, 2008; Rasinski, Rikli, & Johnston, 2009; Schwanenflugel et al., 2004). However, in Clay and Imlach (1971), the rater subjectively recorded sound features due to lack of technology available at the time of research, and ratings lacked reliability and statistical investigation to support the results discussed. In the research conducted by Rasinski et al. (2009) and Klauda and Guthrie (2008), the measure of prosody used was not adequately supported by reliable data. In each of the above-mentioned studies, prosody was assessed spectrographically using speech software, which is not a practical measure for classroom use.

Some have concluded that an emphasis on expressiveness during reading instruction improves rate and comprehension as a result. Martinez, Roser, and Strecker (1998/1999) reported that students made twice the gains in reading rate by using a specific program focusing on prosody. Others have documented phenomenal rate gains when expression and meaning were emphasized instead of speed (e.g., Griffith & Rasinski, 2004; Rasinski & Stevenson, 2005). However, while these studies offer fascinating suggestions regarding prosody and its relationship with reading rate and comprehension, either no specific data were collected on prosodic skills or, again, reliable measures of prosody were not established. Therefore, correlations between prosodic skill, and rate and comprehension cannot be accurately determined.

Text difficulty has appeared to contribute to children's use of prosodic skills while reading. Benjamin and Schwanenflugel (2010) believed that challenging text is more likely to promote the use of prosody because children need it to comprehend what they read. In this study, prosodic features were analyzed spectrographically. Their findings confirmed the results of Young and Bowers (1995); however, the prosody scale they used lacked evidence of reliability.

In addition to studying the relationship between prosody and comprehension, researchers have explored prosodic suprasegmental features. According to the previously mentioned studies that measured speech features, proficient readers end declarative sentences with a purposeful pitch declination and make short pauses (Clay & Imlach, 1971; Miller & Schwanenflugel, 2006; Schwanenflugel et al., 2004). More specifically, good readers make short pauses at internal commas for many types of sentences (Miller & Schwanenflugel, 2006). Miller and Schwanenflugel also found that following yes-no questions, adept readers demonstrate considerable pitch rises. This study also suggested that different prosodic features might be specifically connected to different features of the reading process. For example, the researchers concluded that long pauses might indicate that a reader is having trouble decoding, and that dramatic pitch changes at the end of sentences may indicate a reader's good comprehension skills. In a longitudinal study, Miller and Schwanenflugel (2008) found an association between readers who use appropriate pitch and readers who have fewer pausal intrusions and good comprehension skills. Spectrographic analysis of speech features can provide researchers with dependable findings; however, as mentioned before, it's not practical for classroom use.

In the above-mentioned studies, some researchers recorded student readings and spectrographically analyzed prosodic skills using speech software. While spectrographic analysis of speech features can provide researchers with valuable findings, educators still need an efficient and reliable measure of prosody to completely assess reading fluency. Other researchers have measured prosody with tools lacking reliability evidence.

1.2.2 Prosody measures. Documentation of effective use of oral reading fluency measures has been thorough. What is not as well documented is the use of reliable measures when determining the rate of growth, or correlations to comprehension, in prosody skills. Some examples of current measures include NAEP's Oral Reading Fluency Scale (NCES, 2005) and Zutell and Rasinski's Multidimensional Fluency Scale (Zutell & Rasinski, 1991). Unfortunately, in the case of most measures of prosody, publishers reported very limited reliability data. According to the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999), "When subjective judgment enters into test scoring, evidence should be provided on both interrater consistency in scoring and within-examinee consistency over repeated measurements" (p. 33).

1.2.3 *Zutell and Rasinski's fluency scales.* Allington (1983) created one of the earliest measures of prosody found in the literature. This scale examined phrasing, beginning with "word-by-word" reading with attention to punctuation, semantic, and syntactic cues as well as approximation of "normal speech" (p. 559). Zutell and Rasinski (1991) developed prosody measurement scales originally adapted from earlier measures. They developed uni-dimensional prosody scales, which included appropriate phrasing and intonation as well as a word-per-minute (pacing) component. Zutell and Rasinski also reported impressive average group reliability coefficients (.99); however, this was after group rater training. They continued with their research on measurement of prosody and developed a multi-dimensional scale. This scale incorporate phrasing, smoothness, and pace (with no words-per-minute measure).

Zutell and Rasinski (1991) cited a 1985 study by Rasinski in which a test-retest reliability of .90 and interrater reliability of .96 for grade 3 and .98 for grade 5 were accounted for. However, it is difficult to discern which scale Rasinski used during that particular study. Rasinski, Rikli, and Johnston (2009) also used a version of the Multidimensional Fluency Scale, which they refer to as the Multi-Dimensional Fluency Scoring Guide (MFSG). This instrument was tested for reliability by two of the authors. They claimed a consensus estimate of 81% for two raters, and reliability of 94% for exact or adjacent matches, but they provided no further information regarding what statistical analysis was conducted to arrive at the reported figure.

The Zutell and Rasinski scales measure prosodic reading on "practiced" reading passages. "Practiced" indicates that the students had practiced reading the passages prior to the reading session that was assessed. When calculating reliability on a 4- or 6-point scale, one must be mindful that ratings that fall within one point of each other might be considered agreement simply due to chance. The probability that ratings will fall with plus or minus one on a 6-point scale is very high, simply due to chance. Raters who used the Zutell and Rasinski scales also received extensive training on how to use the assessments. In addition, according to Stemler (2004), interrater reliability must be established for each new study, even if an instrument is supported by previously demonstrated reliability analysis.

1.2.4 ***NAEP Oral Reading Fluency Scale.*** Another uni-dimensional scale was developed by NCES (2005). This scale ranges from a rating of 1 to 4 (Levels 1 and 2 = *nonfluent*; Levels 3 and 4 = *fluent*). Level 4 indicates meaningful phrasing, expression, and appropriate interpretation of author's syntax. In contrast, Level 1 indicates word-by-word reading and inappropriate interpretation of syntax (p. 28). Using a uni-dimensional scale limits score variability. This, in turn, makes it difficult to establish reliability or measure incremental changes in prosodic reading skill. Again, reading was rated using this scale after students had practiced the passage three times. In addition, the passage included illustrations, which do not isolate prosodic skills from non-textual clues.

1.2.5 ***Klauda and Guthrie's scale.*** Klauda and Guthrie (2008) added to previously designed scales to create a measure as a part of a larger study. Their scale evaluated reading on five dimensions using a scale ranging from 1 = *very weak* to 4 = *very strong*. Klauda and Guthrie found that the scorings of the three judges revealed a median correlation of .70. Even on a collapsed scale to "further examine interrater reliability," they were only able to reach 79% agreement between two judges (p. 314).

Other scholars have developed measurement rubrics, also with limited reliability coefficients reported for ratings obtained from those scales. For example, the Reading Teacher Checklist (Hudson, Lane, & Pullen, 2005) includes phrasing and tone as well as syntactic tone. Unfortunately, in the case of all of the referenced rubrics, publishers reported very limited reliability data.

1.3 Implications for Research

Fluency is defined as the rate, accuracy, and prosody of one's reading. Reliable measures of rate and accuracy are readily available, but there is a relative dearth of measures of prosody documented in the research. Therefore, developing reliability on a prosody measure will assist educators when measuring a student's ability to read fluently in a classroom setting. The development of a reliable prosody measure is also needed for further research on prosody—specifically the relationship between prosody and comprehension.

2. Method

2.1 Assessment Description

Our focus in this study was to develop a multi-dimensional prosody rubric. We used the six-step procedure recommended by Gall, Gall, and Borg (2010): (a) define the construct to be measured, (b) define the appropriateness of the construct and measure for a population, (c) review previously developed measures and justify the need for a new measure, (d) develop a prototype, (e) conduct a field test and collect adequate data, and (f) revise the measure based on results. To that end, we reviewed existing prosody assessments and refined items so they more closely correspond to the purpose of this rubric. Each of these prosody measurements has been discussed in detail earlier. The rubric we have devised, known as the WOW Prosody Rubric, uses a 4-point Likert-type scale measuring six dimensions of prosody identified in the literature. These dimensions can be divided into two skill categories: Phrasing and Expression.

Phrasing includes the dimensions of Smoothness and Punctuation. Smoothness is operationally defined across four levels. Level 1 Smoothness includes the following behaviors: "word-by-word reading; frequently extended pauses, hesitations, false starts, sound-outs, repetitions, and/or multiple attempts; struggles to pronounce most words; and does not sound like natural language." Level 4 Smoothness characteristics include the following behaviors: "longer than three-word phrases throughout the entire reading; generally smooth reading with some breaks, but word and structure difficulties are resolved quickly, usually through self-correction; and *text sounds like natural language throughout the entire reading*." Additional definitions are provided for Levels 2 and 3 of Smoothness (see Appendix). The second dimension of Punctuation is also operationally defined across four levels. Level 1 Punctuation is defined as the behavior of "no pauses for punctuation." Level 4 Punctuation is defined as the behavior of "[p]auses at sentences and commas throughout the entire reading."

The second category, Expression, comprises four dimensions: (a) Vocal Emphasis, (b) Inflection, (c) Intonation, and (d) Characterization/Voice. Vocal Emphasis is operationally defined across four levels. Level 1 Vocal Emphasis is characterized as "[n]o change in emphasis at any point in the reading" while Level 4 Vocal Emphasis is characterized as "[a]ppropriate emphasis throughout the entire reading." Inflection is operationally defined across four levels ranging from Level 1 Inflection, characterized as "[n]o change in inflection," to Level 4 Inflection, "[a]ppropriate inflection through the entire reading."

Intonation is operationally defined across four levels ranging from Level 1 Intonation, "[m]onotone," to Level 4 Intonation, "[a]ppropriate intonation throughout the entire reading."

Finally, Characterization/Voice is also operationally defined across four levels ranging from Level 1 Characterization/Voice, defined as "[c]omplete lack of character differentiation," to Level 4 Characterization/Voice, defined as "appropriate character differentiation throughout the entire reading."

2.2 Assessment Development

For the purpose of this study, the oral reading prosody construct involves the ability to read passages of a known difficulty level with expression (discussed more fully above) and is a subset of the broader construct of oral reading fluency.

Prior research has demonstrated the effectiveness of repeated oral reading as an intervention for struggling readers in the development of oral reading fluency (National Reading Panel, 2000; Rasinski & Stevenson, 2005) and specifically in the development of prosodic skills (Kuhn, 2004; Martinez, Roser, & Strecker, 1998/1999). Given this intervention model, measuring prosodic skills has focused on orally reading *practiced* passages. Extant approaches to prosody measurement have also focused on passages that are paired with illustrations or drawings. To isolate prosodic skill from comprehension and memory functions associated with repeated readings and picture-cues, this study uses passages that are not combined with illustrations and have not been practiced prior to reading during the assessment procedure.

The monitoring of growth in prosodic skill, particularly in a response-to-intervention paradigm, is best accomplished by maintaining a constant task demand throughout successive administrations of the assessment. By using the multiple grade-level oral reading fluency (ORF) probes developed for the DIBELS (Dynamic Measurement Group, 2009) that are available through multiple sources (e.g., AIMSweb [NCS Pearson, Inc., 2011]), the established psychometric properties of the probes were used. Additionally, because these probes are already in use, the addition of the prosody rubric to the benchmark and progress-monitoring efforts of teachers will simply add another important dimension to periodic assessment and will enhance the validity of the curriculum-based measurement (CBM) procedures.

2.3 Data Collection

2.3.1 Participants. Two participant groups were involved in this study: readers and raters. The readers were elementary school students in grades one through four (age 6.0 years to 10.0 years). Students were recruited from one elementary school in central Washington State. Letters of support from the district, as well as parental-consent letters, were obtained prior to conducting research. According to the OSPI Washington State Report Card (Office of Superintendent of Public Instruction, n.d.), this school's population had these demographics at the time of research: 53.3% of students were male and 46.7% were female; 83.1% were Hispanic, 15% were white, 0.8% were Black, and 0.3% were American Indian/Alaskan Native. In addition, 88.1% were eligible for free/reduced-priced meals, 7.8% received Special Education services, 1.4% received services on Section 504 plans, 0% were classified as migrant, 44.7% were classified as Transitional Bilingual, and 0% were classified as receiving Foster Care services. Unique health requirements beyond the ability to communicate orally were not a consideration for the purpose of this research study. Children with disabilities that do not impact oral communication were included in the solicitation pool. Students were solicited from all ethnic groups and across a spectrum of reading ability as initially identified by teachers. English Language Learners were also solicited for participation. Approximately 100 students were randomly chosen from a solicitation pool of participants. Students were brought to an empty classroom by whole class, and then individually videotaped. Students were timed while they read DIBELS passages aloud. Only 30 video clips of student readings were selected due to video quality issues.

The second group of participants for this study was the raters. This group included 82 teacher candidates enrolled in one of the following undergraduate education courses at Eastern Washington University: Foundations of Assessment; Introduction to Elementary Reading; or Elementary Reading Methods, Management, and Assessment. These introductory courses were chosen purposefully so that the raters would likely have little knowledge of prosody. This information was intended to determine how much training might be needed to administer the rubric, if needed for further research.

2.3.2 Sampling procedures. Convenience sampling was used for both participant groups in this study. Readers were solicited for participation via letters of invitation sent to parents. Consent was obtained from parents, as was student assent. Consent forms were prepared for the signature of parents and students in both Spanish and English, and translators were available for other languages prior to consent. Raters were solicited for participation via course instructor suggestion. Consent forms were also prepared for rater signature.

2.3.3 Administration procedures. Participating readers were instructed to read for 1 minute from one ORF probe at their instructional level. Each participating student was instructed to read the ORF passage using "their best reading." Their oral readings were videotaped by graduate students, and coded for identification. Edited versions were prepared in which 30 of the original students were shown reading for 1 minute.

Multiple raters rated these 30 edited, videotaped student readings. By using videotaped readings, multiple raters could rate each student's reading prosody using the WOW Prosody Rubric without the necessity for raters to be present during the student's oral reading. Videotapes also allowed for multiple rater administrations. Administration 1 had 45 raters; Administration 2 had 37 raters. A standard script was read for each administration.

Raters also received copies of rubrics for each reader to be rated, a list of passage titles read by each reader, and copies of the text passages being read by the students on the videotapes. The purpose of having text copies available to raters was to compensate for the sound quality of the video. Raters were also given a written survey at the end of session that asked, "Was the sound quality of the video adequate enough to identify a rating?" Raters were to circle either Yes or No. Raters received no training prior to the research administration.

2.4 Data Analysis

2.4.1 Date preparation. Data were coded with rater identification numbers and administration numbers, and entered into SPSS Version 19. Each of the six dimensions rated was entered into individual data sets containing 82 columns of raters by 30 rows of readers.

2.4.2 Descriptive statistics. Intraclass correlations are used to analyze consistency when evaluating multiple raters on ordered category scales (Bresciani et al., 2009). Consistency estimates were used in this study to assess interrater reliability, as consistency estimates evaluate whether judges consistently employ a scoring rubric, and do not depend on consensus or rater training (Stemler, 2004). Statistical analysis of the data was done using a Spearman rho, one form of intraclass consistency correlations. This analysis is appropriate for this study's data set because a Spearman rho is a nonparametric test that will evaluate ordinal data (Harris, 1998).

Eight random pairs of raters were selected for analysis using a random-number generator. These eight pairs of raters were evaluated using a Spearman rho for each of the six dimensions of the rubric. Raters with missing scores on any dimension for any reader were excluded from the analysis, as a Spearman rho requires each case to be judged by each rater (Stemler, 2004). A Spearman rho correlation was provided for each pair of raters on each dimension. The range, mean, and median were calculated for correlations for each dimension.

3. Results

Correlation coefficients for each pair of random raters are shown in Table 1. Each correlation for the dimension of smoothness was statistically significant, where correlations for remaining dimensions were only statistically significant for half of the pairs. The mean correlations between eight pairs of judges range from 0.2973, which is a low correlation, to 0.5277, which is considered a moderate correlation. For this study, because each correlation coefficient is below 0.7, the measure under investigation lacks convincing interrater reliability.

4. Discussion

4.1 Limitations of the Study

This study has several limitations, including a narrow representation of ethnicities in one of the participant groups. As readers were from a school comprising a majority of Hispanic students, they were overrepresented of in the data set. Readers were also videotaped, which limits the rater's ability to accurately rate each reading due to lack of sound quality. The level of understanding of the raters also limited the study. Raters were enrolled in courses at the beginning of their teacher preparation program, indicating that they had minimal knowledge of reading fluency and prosody.

Had the data set represented raters from a wider range of expertise levels (i.e., teacher candidates closer to graduation, paraprofessionals, and teachers), a more accurate account of how much training is needed for raters could have been determined.

4.2 Recommendations

Further research on reading prosody depends on the development of a reliable measure. Additional research that represents a wider range of ethnic groups and in which readers will be assessed in person is needed to establish reliability on a specific prosody measure. A wider range of expertise levels of raters in the data set would provide information to determine at what level of knowledge raters can reliably use the measure.

To increase the functionality of the WOW Prosody Rubric without the need for rater training, we recommend that the rubric be refined (Gall, Gall, & Borg, 2010). Creating more specific descriptions of rubric criteria is one way to improve rater consistency (Bresciani et al., 2009), which can be easily accomplished through the addition of descriptors to better define each prosodic dimension. For example, for raters who can differentiate between Inflection and Intonation, a descriptor of "pitch" can be added to describe Inflection, and "rhythm, tone" can be added to describe Intonation. Under the dimension of Characterization/Voice, an indication of appropriate voice for a specific text not including dialogue would prevent raters from neglecting to rate a reader on this specific dimension. This misperception was evident in the data by missing scores for this dimension.

Since the adoption of the Common Core State Standards (CCSS) by most states, it is imperative that schools and districts target professional development for their instructional staffs. Because the CCSS focuses on fluency—including the elements of prosody—in the foundational skills of reading for grades 1 through 5, we believe training teachers to use the WOW Prosody Rubric—including more refined iterations of it as they are developed—would greatly increase instructional effectiveness in elementary English Language Arts. If students are to succeed on high-stakes measures, all three components of fluency be addressed with professional development as well as in teacher preparation programs. Valid and reliable assessment instruments to assess fluency must be either identified or developed that adequately address this underrepresented component.

4.3 Conclusions

Fluency, defined as the ability to read at a reasonable rate with accuracy and proper expression, is a critical component of proficient reading. Measures of oral reading fluency (rate and accuracy) that are both valid and reliable have been developed and used effectively. However, prosody has been disjointed from its counterparts in current measures of oral reading fluency. Available prosody measures are not supported by reliability data. Without a reliable and valid prosody measure, reading fluency rate of growth cannot be accurately measured. Documentation of reliability on a measure of prosody is fundamental for future research on reading fluency, as well as for functional classroom assessment use in light of the CCSS.

5. References

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

Allington, R. L. (1983). Fluency: The neglected reading goal. *Reading Teacher, 36*(6), 556-561.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Armbruster, C. C., Lehr, F., & Osborn, J. (2003). *Put reading first: The research building blocks for teaching children to read* (2nd ed.). Washington, DC: Partnership for Reading, a collaborative effort of the national Institute for Literacy, the National Institute of Child Health and Human Development, and the U.S. Department of Education. Retrieved from http://www.nifl.gov/nifl/partnership forreading/publications/PRFbooklet.pdf

Benjamin, R. G., & Schwanenflugel, P. J. (2010). Text complexity and oral reading prosody in young readers. *Reading Research Quarterly, 45*(4), 388-404. doi: 10.1598/RRQ.45.4.2

Bresciani, M. J., Oakleaf, M., Kolkhorst, C. N., Barlow, J., Duncan, K., & Hickmott, J. (2009). Examining design and interrater reliability of a rubric measuring research quality across multiple disciplines. *Practical Assessment, Research & Evaluation, 14*(12).

Cassidy, J., & Grote-Garcia, S. (2013). Results of the 2013 What's Hot, What's Not Literacy Survey. *Reading Today, 30*(1), 9-12.

Cassidy, J., Valdez, C. M., & Garrett, S. D. (2010). Literacy trends and issues: A look at the five pillars and the cement that supports them. *Reading Teacher, 63*(8), 644-655.

Clark, V. I. P., & Creswell, J. W. (2010). *Understanding research: A consumer's guide*. Boston, MA: Merrill.

Clay, M. M., & Imlach, R. H. (1971). Juncture, pitch, and stress as reading behavior variables. *Journal of Verbal Learning and Verbal Behavior, 10*(2), 133-139. doi: 10.1016/S0022-5371(71)80004-X

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219-232.

Dowhower, S. L. (1991). Speaking of prosody: Fluency's unattended bedfellow. *Theory into Practice, 30*(3), 165.

Dynamic Measurement Group. (2009). Dynamic measurement group: Supporting school success one step at a time. Retrieved from http://dibels.org

Erekson, J. A. (2010). Prosody and interpretation. *Reading Horizons, 50*(2), 80-98.

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*(3), 239-256.

Gall, M. D., Gall, J. P., & Borg, W. R. (2010). *Applying educational research* (6th ed.). Boston, MA: Pearson Education, Inc.

Gibson, E. J., & Levin, H. (1975). *The psychology of reading*. Cambridge, MA: Massachusetts Institute of Technology Press.

Griffith, L. W., & Rasinski, T. V. (2004). A focus on fluency: How one teacher incorporated fluency with her reading curriculum. *Reading Teacher, 58*(2), 126-137.

Harris, M. B. (1998). *Basic statistics for behavioral science research*. Needham Heights, MA: Allyn & Bacon.

Hasbrouck, J., & Tindal, G. (1992). Curriculum-based oral reading fluency norms for students in grades 2 through 5. *Teaching Exceptional Children, 24*, 41-44.

Hudson, R. F., Lane, H. B., & Pullen, P. C. (2005). Reading fluency assessment and instruction: What, why, and how? *Reading Teacher, 58*(8), 702-714. doi: 10.1598/RT.58.8.1

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. (2003a). Accuracy and fluency in list and context reading of skilled and RD groups: Absolute and relative performance levels. *Learning Disabilities Research & Practice, 18*(4), 237-245.

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

Klauda, S. L., & Guthrie, J. T. (2008). Relationships of three components of reading fluency to reading comprehension. *Journal of Educational Psychology, 100*(2), 310-321.

Kuhn, M. R. (2004). Helping students become accurate, expressive readers: Fluency instruction for small groups. *International Reading Association, 58*(4), 338-344. Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology, 95*(1), 3-21.

Kuhn, M. R., Schwanenflugel, P. J., Meisinger, E. B., Levy, B. A., & Rasinski, T. V. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly, 45*(2), 230-251. Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology, 95*(1), 3-21.

Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology, 95*(1), 3-21.

Martinez, M., Roser, N. L., & Strecker, S. (1998/1999). "I never thought I could be a star": A readers theatre ticket to fluency. *Reading Teacher, 52*(4), 326.

Miller, J., & Schwanenflugel, P. J. (2006). Prosody of syntactically complex sentences in the oral reading of young children. *Journal of Educational Psychology, 98*(4), 839-843.

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.

National Center for Education Statistics. (2005*). Fourth-grade students reading aloud: NAEP 2002 special study of oral reading*. (NCES Publication No. 2006-469). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

National Center for Education Statistics. (2011). *Reading 2011: National assessment of educational progress at grades 4 and 8*. (NCES Publication No. 2012-457). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

National Reading Panel. (2000). *Report of the National Reading Panel: Teaching children to read. Report of the subgroups*. (NJH Publication No. 00-4769). Washington, DC: U.S. Department of Health and Human Services, National Institutes of Health, National Institute of Child Health and Human Development.

NCS Pearson, Inc. (2011). AIMSweb. Retrieved from http://www.aimsweb.com

Office of Superintendent of Public Instruction. (n.d.). Washington state report card. Retrieved from http://reportcard.ospi.k12.wa.us

Pearson Education, Inc. (2012). Assessment and information. GORT oral reading test (4th ed.). Retrieved from http://pearsonassessments.com

Pinnell, G. S., Pikulski, J. J., Wixon, K. K., Campbell, J. R., Gough, P. B., & Beatty, A. S. (1995). *Listening to children read aloud*. (NCES Publication No. 95-762). Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education.

Rasinski, T. (2004). Creating fluent readers. *Educational Leadership, 61*(6), 46-51.

Rasinski, T. (2006). Reading fluency instruction: Moving beyond accuracy, automaticity, and prosody. *Reading Teacher, 59*(7), 704-706. doi: 10.1598/RT.59.7.10

Rasinski, T., Rikli, A., & Johnston, S. (2009). Reading fluency: More than automaticity? More than a concern for the primary grades? *Literacy Research & Instruction, 48*(4), 350-361.

Rasinski, T., & Stevenson, B. (2005). The effects of fast start reading: A fluency-based home involvement reading program, on the reading achievement of beginning readers. *Reading Psychology, 26*(2), 109-125. doi: 10.1080/02702710590930483

Schreiber, P. A. (1991). Understanding prosody's role in reading acquisition. *Theory into Practice, 30*(3), 158.

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. R., Good, R. H., Knutson, N., Tilly, W. D., & Collins, V. L. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21*, 459-479.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4).

Young, A., & Bowers, P. G. (1995). Individual difference and text difficulty determinants of reading fluency and expressiveness. *Journal of Experimental Child Psychology, 60*(3), 428-454.

Zutell, J., & Rasinski, T. V. (1991). Training teachers to attend to their students' oral reading fluency. *Theory into Practice, 30*(3), 211-217.

Appendix

Table 1—WOW Prosody Rubric

WOW Prosody Rubric					
Prosody Qualities	Level	Level 1 Characteristics	Level 2 Characteristics	Level 3 Characteristics	Level 4 Characteristics
Phrasing					
Smoothness		Word-by-word reading; frequent extended pauses, hesitations, false starts, sound-outs, repetitions and/or multiple attempts; struggles to pronounce most words; *does not sound like natural language*	Occasional two- and three-word phrasing; several "rough spots" in text where extended pauses, hesitations, etc. are more frequent and disruptive; *voice begins to make text sound like natural language* in some areas of the reading; focus is still on correct punctuation of words	Several three-word or longer phrases; occasional breaks in smoothness caused by difficulties with specific words and/or structure; *text sounds like natural language for most of the reading*	Longer than three-word phrases throughout the entire reading; generally smooth reading with some breaks, but word-structure difficulties are resolved quickly, usually through self-correction; *text sounds like natural language throughout the entire reading*
Punctuation		No pauses for punctuation	Pauses at sentences *most of the time*; no pauses for commas	Pauses at sentences *all of the time*; pauses at commas *most of the time*	Pauses at *sentences and commas throughout the entire reading*
Expression					
Vocal Emphasis		No change in emphasis	Appropriate emphasis *some of the time*	Appropriate emphasis *most of the time*	Appropriate emphasis *throughout the entire reading*
Inflection		No change in inflection	Appropriate inflection *some of the time*	Appropriate inflection *most of the time*	Appropriate inflection *throughout the entire reading*
Intonation		Monotone	Appropriate intonation *some of the time*	Appropriate intonation *most of the time*	Appropriate intonation *throughout the entire reading*
Characterization/ Voice		Complete lack of character differentiation	*Some indication* of character differentiation	Character differentiation *most of the time*	Appropriate character differentiation *throughout the entire reading*
Total					

Appendix

Table 2 – Spearman Rho Coefficients for Paired Readers

Table 1. Spearman rho Coefficients for Paired Raters

Dimension	Pair 1	Pair 2	Pair 3	Pair 4	Pair 5	Pair 6	Pair 7	Pair 8
Smoothness	0.373*	0.425*	0.424*	0.674*	0.685**	0.627**	0.634**	0.38*
Punctuation	0.329	0.233	0.380*	0.320	0.416*	0.138	0.484**	0.356
Emphasis	0.437*	0.346	0.054	0.581**	0.149	0.332	0.451*	0.429*
Inflection	0.255	0.407*	0.033	0.408*	0.325	0.308	0.327	0.429*
Intonation	0.226	0.179	0.406	0.461*	0.600**	0.249	0.353	0.365*
Characterization/ Voice	0.171	0.322	0.449*	0.472**	0.356	0.330	0.162	0.117

Note. * $p < .05$. ** $p < .01$.