DEVELOPMENT AND VALIDATION OF THE COMPREHENSIVE ORAL READING FLUENCY SCALE

by

REBEKAH ANNE BENJAMIN

(Under the Direction of Paula J. Schwanenflugel)

ABSTRACT

Continued technological advances as well as a renewed interest in the construct of reading fluency have made scientific studies of fluency's component parts more accessible over the past several decades. While it is now generally understood that fluent readers read with appropriate rate, accuracy, and expression (Kuhn, Schwanenflugel, & Meisinger, 2010), very little has been done in changing the assessment of children's oral reading fluency from a system based on automaticity alone to a system in which all dimensions of fluency are measured. Some tools have been developed to measure fluency as a complex construct, but little psychometric information is available to demonstrate their reliability and validity. The present studies detail the development and the testing of the Comprehensive Oral Reading Fluency Scale, a new scale grounded in spectrographic measurements of oral reading prosody, which allows users to measure the multiple components of oral reading fluency. Validation of the scale is based on Kane's (1992) argument-based validation framework. The scale was developed and tested with second and third grade children, and interrater reliability was analyzed using reading experts as raters. The relationships between scale ratings and spectrographic measures of oral reading prosody as well as between ratings and traditional measures of reading skill served as evidence of the scale's validity. In general, the scale performed well as a tool for providing users with both general and specific information about children's oral reading fluency.

INDEX WORDS: Reading fluency, Prosody, Oral reading, Assessment, Rating scale, Elementary school, Reading, Comprehension, Spectrographic measurement

DEVELOPMENT AND VALIDATION OF THE COMPREHENSIVE ORAL READING FLUENCY SCALE

by

REBEKAH ANNE BENJAMIN

A.B., Indiana Wesleyan University, 2004

M.A., The University of Georgia, 2009

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2012

DEVELOPMENT AND VALIDATION OF THE COMPREHENSIVE ORAL READING FLUENCY SCALE

by

REBEKAH ANNE BENJAMIN

Major Professor:	Paula J. Schwanenflugel
Committee:	Stacey Neuharth-Pritchett
	Nancy F. Knapp
	Shawn Glynn

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2012

DEDICATION

To my parents, who taught me how to work. To Nate, who has loved and supported me throughout these long years of graduate school. To Darcy, who helped me laugh through the year I wrote this dissertation.

ACKNOWLEDGEMENTS

I want to thank Paula Schwanenflugel for her unrelenting support and motivation while I’ve been her student these past four years. I got lucky to have her for an advisor and couldn’t have asked or wished for better. My committee has been fabulous—Stacey, Nancy, and Shawn—and I am blessed to have developed a positive and supportive relationship with each member. Thank you to Karen Samuelsen and Debbie Bandalos for excellent statistical instruction in my courses and for occasional but helpful advice on statistics and measurement issues while I conducted this and other research. Stephanie Lai and Adam Reed volunteered many hours of their time to help me collect data for this research while I was in my third trimester of pregnancy. I am forever in their debt! Finally, I honestly would not have made it this far without my husband, Nate. During my many moments of doubt and willingness to compromise, he kept reminding me that I could and would accomplish this goal…or I would regret it for the rest of my life. Those words were always what I needed to press on.

TABLE OF CONTENTS

Page

APPENDICES

LIST OF TABLES

Page

LIST OF FIGURES

Page

Chapter 1: Literature Review

Oral reading fluency has shifted in the past decades from a neglected reading skill (Allington, 1983; Dowhower, 1991) to a heavily assessed, studied, and controversial facet of overall reading ability. In fact, the appropriate assessment of reading fluency has changed considerably over the past 50 years (Fuchs, Fuchs, Hosp, & Jenkins, 2001), and while fluency is certainly more researched and taught in teacher preparation programs than previously, widely varying methods of assessing oral reading fluency still exist and compete for use in schools and research (Kuhn, Schwanenflugel, & Meisinger, 2010). Unfortunately, ease of use can win out over thoroughness, in which case the validity of inferences made from assessments (and many inferences are often made from these assessments) may be compromised. However, in classrooms—where time is very valuable—assessments that provide quick and reliable results may be the only option.

It is critical to determine, then, whether current assessments of oral reading fluency line up with current construct definitions of oral reading fluency. In classrooms, where one holistic assessment of oral reading fluency can certainly save time over multiple individual component assessments, are available holistic assessments appropriate for measuring what they purport to measure? Do they line up with standards of reliability and validity? Finally, at a time when research regarding prosody's function in fluency is rapidly expanding, new measurement tools must be developed to reflect this increased understanding.

The following review begins with a discussion of oral reading fluency's definition, its components, and its connection to reading comprehension. Traditional measures of oral reading fluency are described as well as methods that have been used, specifically, for measuring prosody—the most recently focused on component of oral reading fluency. Three popular oral reading fluency scales are analyzed and critiqued in depth and compared in terms of their application of currently defined fluency components and their psychometric properties. Finally, the framework for developing a new comprehensive oral reading fluency assessment is presented and the purpose of the subsequent new scale validation study is discussed.

Oral Reading Fluency

Definition of the construct. Oral reading fluency is commonly used as a simple but powerful indicator of comprehension in young children, but recently it has gained attention as a complex construct in its own right (e.g., Deeney, 2010; Hudson, Pullen, Lane, & Torgesen, 2009; Kuhn et al., 2010; Kuhn & Stahl, 2003; Wolf & Katzir-Cohen, 2001). It is now widely agreed that fluent oral reading is comprised of multiple components, some of which are easy to measure—reading rate and accuracy—and some which are more complicated—expression, prosody, flow, or naturalness.

Recent definitions of reading fluency differ in some details but appear to largely agree on fluency as a multi-dimensional construct (see Kuhn et al., 2010, for a review of fluency definitions). Fluent oral reading has been defined as "accurate reading of connected text at a conversational rate with appropriate prosody or expression" (Hudson, Mercer, & Lane, 2000, as cited in Hudson, Lane & Pullen, 2005, p. 702). The National Reading Panel (National Institute of Child Health and Human Development [NICHHD],

2000) also defined fluent readers as those who can read accurately, at an appropriate rate, and with proper expression. Finally, researchers conducting a recent review analyze multiple definitions of fluency and conclude by developing a theoretically based and operationally useful definition:

> Fluency combines accuracy, automaticity, and oral reading prosody, which, taken together, facilitate the reader's construction of meaning. It is demonstrated during oral reading through ease of word recognition, appropriate pacing, phrasing, and intonation. It is a factor in both oral and silent reading that can limit or support comprehension. (Kuhn et al., 2010, p. 240)

These recent definitions, and others, describe reading fluency largely in terms of three major components: automaticity, accuracy or word recognition, and reading prosody (cf. Daane, Campbell, Grigg, Goodman, & Oranje, 2005; Samuels, 2006).

Automaticity—often measured as reading rate—is actually the result of numerous skills working together in a way that allows the reader to achieve effortless decoding of text (see LaBerge & Samuels, 1974; Logan, 1988; 1997; Perfetti, 1985; Samuels, 1994). Wolf and Katzir-Cohen (2001) list many lower level factors as contributors to rapid reading ability, including skills related to attention, visual perception, orthographic identification, phonological representation, decoding, and word identification among others.

Accuracy is a skill that is sometimes left out of fluency measures (e.g., Pinnell et al., 1995), but is typically included with reading rate in simple assessments of readers' words correct per minute (WCPM), used often in schools, standardized tests, and research. 95% is typically considered a standard for the level of accuracy a reader should achieve if reading texts within his or her instructional level (McKenna & Stahl, 2009). Because accuracy in sight-word recognition and decoding is directly connected with the ability to read connected texts quickly (e.g., Hudson et al., 2009; Wolf & Katzir-Cohen., 2001), it is difficult to separate accuracy from automaticity. However, a recent intervention study found that while having children practice improving accuracy versus accuracy plus reading rate did not lead to different results in comprehension, students who practiced accuracy alone did not make as significant improvements in automaticity as students who practiced rate as well (Hudson, Isakson, Richman, Lane, & Arriaza-Allen, 2011).

Prosody and its measurement are discussed more fully in following sections. As a component of oral reading fluency prosody is the relative new-comer. It is well known, though, that children who read with good fluency also tend to read with appropriate pitch variation, pause structure, and stress (e.g., Benjamin & Schwanenflugel, 2010; Clay & Imlach, 1971; Daane et al., 2005; Miller & Schwanenflugel, 2006; 2008; NICHHD, 2000; Pinnell et al., 1995; Schwanenflugel, Hamilton, Kuhn, Wisenbaker, & Stahl, 2004). Various methods for measuring expressive skills have been developed. Some of these focus on measuring aspects of reading prosody, such as rhythmic sensitivity (e.g., Blumstein & Goodglass, 1972; Whalley & Hansen, 2006) or direct acoustic features like pitch contour, pausing, and stress patterns displayed by children as they read (e.g., Benjamin & Schwanenflugel, 2010; Dowhower, 1987; Koriat, Greenberg, & Kreiner, 2002; Miller & Schwanenflugel, 2006; 2008; Ravid & Mashraki, 2007; Schwanenflugel et al., 2004). Others use rating scales that incorporate prosodic features as an aspect of the rating system (e.g., Allington, 1983; Daane et al., 2005; Pinnell et al., 1995; Zutell & Rasinski, 1991).

Kuhn et al.'s (2010) definition moves beyond simple description and further defines how fluency is evidenced in oral reading through word recognition, pacing, phrasing, and intonation. The end goal, of course, is that children can read with understanding, and while the exact nature of fluency's interaction with comprehension is not yet fully understood, theorists have made attempts at developing models which may explain this relationship.

Connection with comprehension (theoretical models of reading). Reading comprehension is a complex construct and is conventionally viewed as the end goal of reading. Because of its complexity and the expertise required to design valid and reliable measures of comprehension, classroom teachers and reading specialists often use fluency measures throughout the school year to assess a child's reading development. Although fluency is only one (albeit necessary) component of reading ability, it is often used as a reliable predictor of reading comprehension, and improvements in a child's oral reading fluency tend to result in comprehension improvements as well (see Fuchs et al., 2001, for a review). Readers who have developed skills in the various components of fluency are better able to construct meaning from a text (Kuhn et al., 2010; Kuhn & Stahl, 2003). Kuhn et al. (2010; see also Kuhn & Stahl, 2003) described two primary perspectives regarding the effect fluency has on comprehension. One perspective focuses on the role of automatic word recognition in fluent reading, and the other focuses on the role of prosody. Each component of fluency may contribute in several ways to reading comprehension.

Automatic word recognition. Although various theories exist which try to account for and explain the development of automatic processes and, specifically, the potential contribution of automatic word recognition to reading comprehension, these theories agree that practice assists in the development of automatic word recognition (LaBerge & Samuels, 1974; Logan, 1988; 1997; Perfetti, 1985; Samuels, 1994). While these theorists disagree about whether automaticity develops through a change in resource distribution (LaBerge & Samuels, 1974; Perfetti, 1985; Samuels, 1994) or a movement from algorithmic processing to a gradual reliance on episodic memory recall (Logan, 1988; 1997), they agree that practice and exposure with attention assists in the development of automaticity, which can greatly affect reading comprehension.

When reading a text, readers must attend to two broad interdependent tasks: determine what words comprise the text, and determine the meaning that the text is trying to convey. Of course, both of these broad tasks are comprised of several smaller tasks (see Logan, 1997; Rumelhart, 1994; Samuels, 1994). Because these tasks are accomplished simultaneously, readers are required to constantly divide their attention between the two major tasks, and a lack of automaticity at a lower level of processing (e.g., letter level or word level) can impede the rate of higher level processing (e.g., sentence level or text level; Logan, 1997; Rumelhart, 1994; Samuels, 1994; Wolf & Katzir-Cohen, 2001). With practice, however, lower level processes become automatic and higher level processes are less disrupted. In fact, numerous studies in automaticity have found that when a process—such as reading—becomes automatic, an individual can then begin attending to another task while still accomplishing the first (see Logan, 1997, for a review). Because a complex interaction between visual processing and semantic and episodic memory activation takes place during reading (Logan, 1997; Samuels, 1994), automaticity may result from a single exposure or it may take several, but each exposure to a word, letter, or phrase, for example, can increase the reader's ability to decode and process text automatically.

Until readers are automatic in their ability to decode words, they are overly reliant on alternative knowledge sources to guide them through the text (e.g., orthographic, semantic, and syntactic information). The reader cannot attend to two non-automatic tasks at once, so comprehension suffers (LaBerge & Sameuls, 1974; Logan, 1997; Rumelhart, 1994; Samuels, 1994). Thus, automatic word recognition allows readers to concentrate on the meaning of the text while word recognition takes place with little or no effort. Repeated reading allows readers to deepen memory traces of learned words, sentences, and texts, and wide reading allows for transfer and increased learning (Logan, 1997). Thus, through exposure and practice readers will decode print automatically allowing for higher level processes to also improve.

Prosody. Kuhn et al. (2010; see also Kuhn & Stahl, 2003) also describe a complementary perspective on fluency's relationship with comprehension—focusing on prosody's impact. A child can read a text with appropriate rate and accuracy, but research shows that skilled readers also incorporate proper phrasing and expression (Benjamin & Schwanenflugel, 2010; Clay & Imlach, 1971; Dowhower, 1987; 1991; Miller & Schwanenflugel, 2006; 2008; Schreiber, 1987; 1991). Both Logan (1997) and Samuels (1994; LaBerge & Samuels, 1974) claim that once readers have learned words, they are able to both better comprehend text and appropriately organize their prosody. But prosody might not simply be a passive outcome of automatic decoding skills. Assistance in phrasing has been found to help readers comprehend text (Cromer, 1970; O'Shea & Sindelar, 1983) though specific prosodic modeling has not yet been found to have an impact (Young, Bowers, & MacKinnon, 1996).

The hypothetical impact of this prosodic aspect of fluent reading on comprehension is not fully understood, but some research has supported the idea that prosody serves as a naturally-occurring scaffold for comprehension (Frazier, Carlson, & Clifton, 2006; Swets, Desmet, Hambrick, & Ferreira, 2007). It is well known among speech-communications researchers that prosody in spoken language helps listeners understand what the speaker is trying to communicate (See Cutler, Dahan, & van Donselaar, 1997, for a review), and prosody can be especially helpful—and necessary—when syntax obscures the intended meaning. Of course, many of the cues available to listeners are not available to readers; however, research in spoken communication also points to prosody as a tool for chunking speech in working memory, allowing the listener to recall larger segments of speech than would otherwise be possible (Frazier et al., 2006; Swets et al., 2007). This function of prosody might also be at work when reading texts, and recent research suggests that this scaffolding might help readers comprehend texts when rate and accuracy alone are not sufficient (Benjamin & Schwanenflugel, 2010; Valencia et al., 2010).

Recent research on fluency's relationship with comprehension. Fluency plays a significant role in both predicting and assisting in the construction of meaning from text. The most recent research continues to support the long-held belief that children's fluency can predict both their current as well as future comprehension of text, even when different measures of comprehension are used (e.g., Benjamin & Schwanenflugel, 2010;

Klauda & Guthrie, 2008; Miller & Schwanenflugel, 2008; Nunez, 2009; Pangrac, 2009; Roehrig, Petscher, Nettles, Hudson, & Torgesen, 2008; Suchey, 2009; Valencia et al., 2010).

Using only a measure of WCPM, Miller and Schwanenflugel (2008) significantly predicted concurrent comprehension in third graders using a standard reading comprehension test (r^2 = .268). They also found that simple measures of sight-word reading ability in first and second grade significantly predicted students' future comprehension scores in third grade (r^2 = .289 and .393, respectively). Roehrig et al. (2008) used a limited fluency assessment to predict concurrent comprehension. Third graders' DIBELS oral reading fluency scores significantly predicted comprehension on both a state mandated assessment as well as a national standardized assessment (r^2 = .49 for both assessments).

Benjamin and Schwanenflugel (2010) used measures of sight-word reading ability; and connected text rate, accuracy, and prosody (measured via spectrograph) to predict concurrent comprehension of second graders on a standardized test of reading comprehension. All fluency variables combined to account for 60% of the variance in comprehension scores. Also measuring fluency in a comprehensive way, Valencia et al. (2010) found that second, fourth, and sixth graders' reading rate, accuracy, and prosody (measured by a holistic rubric) together accounted for 34-36% of the variance in concurrent ITBS comprehension scores in a simple path model, though model fit indices were not provided or discussed.

Finally, Klauda and Guthrie (2008) conducted a short-term longitudinal examination of fluency's ability to predict comprehension in fifth graders. They found that word reading speed, phrasing, and expression accounted for 56% of the variance in concurrent reading comprehension scores. When predicting changes in comprehension scores at time point 2 (after 12 weeks), Klauda and Guthrie's time point 1 fluency measure (called syntactic processing) significantly predicted time point 2 comprehension after controlling for time point 1 comprehension. However, while this relationship was significant (possibly due to a fairly large sample size), the R^2 value of fluency's unique contribution was only .004. Fluency, then, can serve as a powerful concurrent predictor of whether or not a child is progressing adequately in their overall reading skills, but the direct longitudinal relationship between fluency and comprehension is still not well established (Lai, Benjamin, Schwanenflugel, & Kuhn, in press).

Traditional Measures of Oral Reading Fluency

Though definitions of fluency illustrate widespread agreement that fluency is comprised of multiple components, many fluency assessments use only the simple measure of words correct per minute (WCPM) to represent the fluency construct. Curriculum Based Measures (CBMs) of oral reading fluency (Deno, 1985), Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002), and fluency components of tests like the Gray Oral Reading Test (GORT; Wiederholt & Bryant, 1992, 2001) and Qualitative Reading Inventory (QRI; Leslie & Caldwell, 2011) largely measure fluency by taking reading rate and reading accuracy into account. Additionally, when educators measure WCPM, students are often not given any comprehension measures based on the passage and are not instructed to read for comprehension, which Samuels (2007) describes as problematic.

These WCPM measures are considered to be reliable and generally valid (cf. Shelton, Altwerger, & Jordan, 2009), but a simple examination of Kuhn et al.'s (2010) definition of fluency as well as some recent research suggests that the WCPM method alone may not be as valid as fluency measures that also take prosody into account (Benjamin & Schwanenflugel, 2010; Miller & Schwanenflugel, 2006; 2008; Valencia et al., 2010). In fact Valencia et al. (2010) found that adding prosody rating scale scores to a WCPM measure increased predictive power, especially at the upper elementary grades, when WCPM begins to steadily decline as an effective predictor of overall reading ability. They also found that using WCPM as a screen for identifying students at risk for reading difficulties resulted in unacceptable numbers of false negatives and false positives—though false negatives, or failing to identify a student who needs assistance, are more disconcerting. Additionally, studies by Benjamin & Schwanenflugel (2010) and Miller and Schwanenflugel (2006; 2008) found that in the primary grades prosody accounts for variance in comprehension scores left unexplained by oral reading rate and accuracy; thus incorporating prosody into fluency would increase the predictive value of fluency measures.

Reading rate and accuracy are simple to measure and are reliable indicators of comprehension, in general (e.g., Deno & Marston, 2006; Fuchs, Fuchs, & Maxwell, 1988). However, while many of these measures are effective in assessing how many words a student can read correctly in a minute, using these simple measures as proxies for more comprehensive measures of reading skill can be problematic. In their discussion of current fluency assessment practices, Kuhn et al. (2010) describe the potential dangers of basing instructional decisions simply on results of WCPM assessment. One problem is that too great a focus on increasing oral reading rate could actually have a negative impact on comprehension (Samuels, 2007), which may be neglected in instruction. These incomplete assessments of fluency are likely a result of the difficulty of finding simple, objective, precise, and psychometrically sound fluency assessments which include prosody.

Reading Prosody and its Measurement

For decades appropriate prosody in oral reading has been found to be a characteristic trait in skilled readers and a trait that is lacking in struggling readers (Benjamin & Schwanenflugel, 2010; Clay & Imlach, 1971; Cowie, Douglas-Cowie, & Wichmann, 2002; Dowhower, 1987; Miller & Schwanenflugel, 2006; 2008; Schwanenflugel et al., 2004). Prosody is a characteristic of spoken language typically measured in terms of loudness, duration, pitch, and pause (Couper-Kuhlen, 1986). Loudness (or indicated spectrographically by signal amplitude) is often modified to place stress on a particular word, phrase, or exclamation; duration can involve rhythm, vowel length, and even the lengthening of an entire word for emphasis; pitch is measured in Hertz (Hz) and is also called intonation or fundamental frequency (F_0); pause is also used for emphasis, to divide an utterance into its major syntactic components and also to signal turn taking in dialogue.

These qualities of spoken language transfer in some respects to oral reading. Of course, the task of the reader (who is required to understand and convey the meaning intended by a writer) is more difficult than the task of the speaker (who is simply required to convey his or her own intended meaning). Readers, however, have several tools at their disposal—at least once they become fairly skilled readers. By incorporating the use of an *inner voice*, readers can phonologically re-code text to make it match up better with speech (Share, 1999). Likewise, punctuation (like commas) can serve as prosodic boundary cues for many readers, at least those who use punctuation appropriately in their own writing (Chafe, 1988; see Steinhauer, 2003 for a review). Prosody, however, is not the same as syntax, and its function in the reading process is not fully understood.

Since standards in measuring reading prosody have not yet been set, researchers have used a variety of methods to measure prosodic skills in children. The following sections discuss and evaluate educational prosody research that has developed through three common methods for measuring prosodic ability: indirect measures (including tests of rhythmic sensitivity and parsing tests), rating scales, and direct (typically spectrographic) measures.

Indirect measures of prosodic understanding. A number of researchers have identified children’s metalinguistic awareness of prosodic features as an indicator of their overall reading skill. Comparing English to Hungarian children (ages 7-11)—using tasks in their native languages—Surányi et al. (2009) examined differences and similarities in children’s prosodic sensitivity that may result from differences in the stress patterns of these two languages. They found that regardless of language, children with dyslexia were less sensitive to stress (specifically, rise time) than children without reading difficulties. Though sensitivity to these acoustic cues correlated with phonological abilities across languages, the English children were more sensitive overall to rise time than Hungarian children; English children also showed greater differences between dyslexic children and their age-matched controls. Authors found that stress sensitivity in English may be “shaped by the need to store metrical patterns that can act as templates for different words" (p. 55). These stress-based measurement methods, then, may be useful in examining children's ability to intuitively grasp the rhythms of their language, and it appears that this ability is related to reading disabilities.

Other sorts of rhythmic sensitivity tests have been used to also examine differences in children without diagnosed reading disabilities. The 'DEEdee' task is used to examine children's ability to differentiate between identical compound nouns and noun phrases using only prosodic cues (Blumstein & Goodglass, 1972). Whalley and Hansen (2006) used this task as well as a lexical rhythmic sensitivity task to examine the relationship between children's prosodic skills and their reading ability. They found that children's performance on the lexical rhythm task predicted their ability to accurately identify words, while performance on the DEEdee task predicted unique variance in reading comprehension. Since the DEEdee task assesses phrase-level prosodic skills and the lexical rhythm task assesses word-level prosodic skills, these measures seem consistent with the way prosodic phrasing works in spoken English—both lexical stress and phrasal stress patterns are important for interpretation. These studies demonstrate that much of what is necessary for interpreting spoken English may also be necessary for developing reading skills.

Rhythm tasks, however, are generally restricted to receptive prosodic awareness. Text parsing (dividing text into various syntactic word group boundaries) requires readers to determine where they should place boundaries in a given text. The task, then, is more interactive than rhythm tasks, but interpretation can be difficult since the child performs the task while reading silently. Researchers conducting an English language study with American children (Kleiman, Winograd, & Humphrey, 1979) found that when 4th grade below-average readers were given a text-parsing task, they performed significantly better when they were able to listen to the text being read aloud (the “prosody condition”) versus without any auditory assistance (the “no prosody condition”). Below-average readers in the prosody condition performed the same on the task as above-average readers. While Kleiman et al. (1979) did not find significant differences among below-average and above-average readers, the overall trend in their results is consistent with more recent work (Young & Bowers, 1995) in which average readers consistently performed better on parsing tasks than below-average readers across various levels of text difficulty. The phrasal knowledge needed to successfully parse sentences is linked with more general reading skills like comprehension as well as measures of fluency. Parsing tests, then, might really be measuring children’s syntactic awareness as well as their ability to comprehend text well enough to segment it into meaningful groups. Presently, though, other means of measuring use of phrase and sentence-level prosody during silent reading are limited.

Direct measures of prosody. Because of the technology and skill required to directly and precisely measure prosodic qualities like pitch, pause, and stress, studies which do this are limited. To closely examine the prosodic features which are characteristic of skilled vs. less skilled readers, it is necessary to have a precise measure of these features. Having this data would not only allow researchers to compare ways in which readers differ prosodically, but would also provide researchers with tools to develop and validate user-friendly assessments that incorporate prosody. Techniques for measuring prosodic features in reading are not standardized, and researchers are still experimenting with the best ways to directly measure prosody (cf.; Benjamin &

Schwanenflugel, 2010; Dowhower, 1987; Koriat et al., 2002; Miller & Schwanenflugel, 2006; Ravid & Mashraki, 2007; Smith, 2004; among others)—probably because there is little consensus about which prosodic features ought to be measured.

One method of directly measuring prosody, using limited technological resources, is to create a prosodic *map* of a text based on expert readings and then simply score children's oral reading according to their relative conformity to the map (Ravid & Mashraki, 2007). This method can be useful for purposes of a single study, but might be too labor intensive for larger studies and may not be appropriate for drawing generalizations of prosodic behavior across several texts.

Other studies in communication sciences and linguistics have used the Tone and Boundary Indices (ToBI) system for analyzing English prosody (e.g., Frazier et al., 2006; see also Zervas, Fakatokais, & Kokkinakis, 2008, for an adapted version for use with Greek prosody), which focuses particularly on pitch accents within speech as they relate to prosodic phrasing. This system has not been widely used in the educational or psychological literature regarding reading, but because of its wide use in disciplines analyzing speech, the ToBI system might be useful for analyzing prosody in children's oral reading. However, the system would likely be too labor intensive for studies incorporating more than a handful of participants, as ToBI labeling typically takes even experienced labelers 100-200 times the actual recording time (Srydal, Hirschberg, McGory, & Beckman, 2001).

Other varied forms of spectrographic measurements are commonly used in education-focused studies for analyzing prosody in oral reading (e.g., Schwanenflugel & Benjamin, 2012; Dowhower, 1987; Koriat et al., 2002; Miller & Schwanenflugel, 2006;

2008; Schwanenflugel et al., 2004). While studies differ widely on specific measures taken from spectrographic software, common themes have emerged. Certain patterns of pause placement and duration, pitch movement, and stress placement within sentences have all been linked with reading skill in children (Benjamin & Schwanenflugel, 2010; Dowhower, 1987; Miller & Schwanenflugel, 2006; 2008; Schwanenflugel et al., 2004) as well as with typical patterns in adults (Koriat et al., 2002). As stated above, though, methods used for measuring these features differ. For example, Dowhower (1987) marked pauses as inappropriately lengthy in children's reading if pauses were greater than 210 milliseconds. Miller and Schwanenflugel (2008) and Benjamin and Schwanenflugel (2010), however, used 100 milliseconds as the cutoff criteria.

Since studies using these methods for precisely measuring prosodic features are still rare in the reading literature, little attention has been drawn to these methodological variations. As researchers continue to gain access to this technology (as well as appropriate training), though, the demand for consistency in both measurement practice and definitions of terms will hopefully increase. If spectrographic measurement is to assist researchers in furthering knowledge of reading processes, then the studies must be easily interpretable.

Ratings as measures of reading prosody. The most commonly used tool for measuring prosody is the rating scale. Numerous scales exist—many designed only for use within single studies—but only a few are consistently found throughout the literature and in classrooms. These include the Allington (1983) scale, various versions of a scale designed by Rasinski and colleagues (Rasinski, 2004; Rasinski, Rikli, & Johnston, 2009; Zutell & Rasinski, 1991), and the National Assessment of Educational Progress (NAEP) scale (Daane et al., 2005; Pinnell et al., 1995). Each scale differs in scope and format, though all scales have been developed with the goal of formally incorporating prosody into fluency measures. Such rating scales succeed in adding this third dimension of fluency—in addition to rate and accuracy—that more traditional measures like the GORT (Weiderholt & Bryant, 2001) have not incorporated. These scales were designed to measure fluency as a whole, not simply prosody, so interpreting the ratings can be complicated. These scales are described and critiqued in detail later in this chapter.

There is still only limited empirical evidence supporting the hypothesis that prosody contributes to comprehension independently of reading rate and accuracy (e.g., Benjamin & Schwanenflugel, 2010; Klauda & Guthrie, 2008; Miller & Schwanenflugel, 2006; 2008; Schwanenflugel et al., 2004; Valencia et al., 2010). In spite of the limited research including prosody in fluency assessments, researchers and educators continue to emphasize the importance of taking prosody into account when judging students' oral readings, and several researchers have provided recommendations for measuring the multiple dimensions of fluency (e.g., Dowhower, 1991; Fuchs et al., 2001; Hudson et al., 2005; Hudson et al., 2009; Kuhn et al., 2010).

Suggestions for Valid Measures of Fluent Reading in the Literature

Over the years several reviews of fluency research have provided researchers and educators with recommendations for measuring oral reading fluency or its various components. Dowhower (1991) did not necessarily provide recommendations for assessment, but did state that six markers appear to be related to prosodic reading and can be examined: pausal intrusions, length of phrases (number of words read between pauses), appropriateness of phrases (does the reader chunk text appropriately, or do phrases cross boundary lines like punctuation marks, split prepositional phrases, or separate nouns from their determiners), phrase-final lengthening, terminal intonation contours, and stress (Clay & Imlach, 1971, found that the best readers stressed one word every 4.7 words, the lowest readers paused nearly between every word and often inserted more than one stress per word). This type of analysis is possible largely with spectrographic measurement methods, the method which Dowhower (1987) used in her own research.

Fuchs et al. (2001) indicate the importance of measuring prosodic features when looking at a child's reading fluency, but—like Dowhower (1991)—underscore the difficulty in reliably and efficiently measuring these features. Thus, their recommendations focus on simpler, more straightforward methods of measurement such as WCPM and miscue analysis. Because of the development and popularization of various fluency rating scales as well as the increased use of CBMs in the classroom, later reviews (Hudson et al., 2005; Hudson et al., 2009; Kuhn et al., 2010) are able to provide more diverse and holistic recommendations for measuring oral reading fluency.

Hudson et al. (2005) recommend methods for measuring each of the three components in their fluency definition: accuracy, rate, and prosody. Teachers and researchers can measure accuracy by counting errors or by conducting miscue analysis; rate should be measured in context and aloud, and WCPM is an ideal method for this. Finally, prosody can be measured using a rating scale like the Multidimensional Fluency Scale (Zutell & Rasinski, 1991) or through a checklist that incorporates appropriate emphasis, appropriate tone, punctuation, appropriate inflection, appropriate vocal tone to illustrate characters' mental states, and appropriate pausing.

In keeping with their earlier review, Hudson et al. (2009) reiterate the importance of measuring children's reading aloud and emphasize the value of using multiple methods for measuring oral reading fluency if one wants to obtain an accurate and valid picture of a child's abilities. They again recommend that educators measure children's prosody, but discuss the difficulty of doing so when so little psychometric information exists for available tools. Kuhn et al. (2010) agree with this criticism of current prosody assessments and state that presently, spectrographic measurement is currently the most precise and valid means for measuring prosody. Unfortunately, this method is far beyond the reach of most practitioners and is incredibly time consuming as well. Thus, rating rubrics or scales appear to be the most likely direction that prosody measurement should take. Unfortunately, however, these scales currently have little psychometric information and do not allow for variations in text difficulty to be taken into account.

Presently, there are numerous studies discussing the pros and cons of assessment methods like WCPM, CBMs, and miscue analysis. However, there is no study available which discusses the psychometric properties and usefulness of some of the most widely used fluency or prosody rating scales. If oral reading fluency is to be measured accurately and completely, then this prosodic component needs to be assessed appropriately. For the sake of current practice and future research, then, the following section describes the available psychometric properties, strengths, and weaknesses of three well-known fluency scales: the scale published by Allington (1983; Allington & Brown, 1979); the scale designed for the NAEP studies (Daane et al., 2005; Pinnell et al., 1995); and the various versions of a scale first published by Zutell and Rasinski (1991; Rasinksi, 2004; Rasinski et al., 2009).

Existing Rating Scale Assessments of Oral Reading Fluency

An assessment can never be completely validated—rather, evidence of its usefulness in making certain types of inferences can be gathered and presented (Crocker & Algina, 1986). Thus, three well-known fluency scales are examined and evaluated based on the following information: 1) description of the scale's content; 2) development of the scale; 3) studies in which the scale has been used; 4) reliability evidence; 5) validity evidence. An overall evaluation is provided for each scale based on its alignment with modern definitions of reading fluency (e.g., Kuhn et al., 2010) and its psychometric properties as evidenced in the literature.

Allington's fluency scale. The Allington Fluency Scale (Allington, 1983; Allington & Brown, 1979) was originally designed as part of a larger reading program entitled *FACT: A Multi-Media Reading Program* (Allington & Brown, 1979) and was adapted from an earlier scale by Mark Aulls (1978). Allington published an adapted version of the scale in a later publication (Allington, 1983). Because the original reading program (Allington & Brown, 1979) is no longer in print and appears to be unavailable through internet resources, the adapted version of the scale is discussed here (see Table 1.1). The scale does not appear to be limited to a specific population of students, but it is evident that the scale is intended for school-aged children, and Allington's 1983 article was published in a journal intended for educators of children up to age 12.

Purpose of test and suggested uses. When Aulls (1978) published his original scale, he included it in an intervention method for students in the middle grades who could decode text but failed to comprehend (see Aulls, 1978, pp. 275-299). While Allington does not explicitly provide instructions for using the scale in his 1983 article,

Table 1.1

Allington Fluency Scale

Reader reads:	**Score**
word by word.	1
primarily word by word with some 2-3 word phrasing.	2
primarily in phrases (2-3 words) but sometimes word by word; sometimes gives phrases inadequate stress in relation to syntax.	3
primarily in phrases with very little word by word reading; sometimes ignores external punctuation; generally reads in a monotone.	4
primarily in phrases, attending to terminal punctuation; some internal punctuation is ignored; expression is not consistently adequate.	5
in phrases, with fluency, using both terminal and internal punctuation; provides appropriate semantic and syntactic emphasis for purposes of dramatization; expression approximates normal speech.	6

Note: From Allington, R. (1983). Fluency: The Neglected Reading Goal. *Reading Teacher*, *36*(6), 556-61.

Meyer and Felton (1999)—referring to Allington and Brown's (1979) *FACT* reading program—state that the scale must be used with two independent raters. Given the necessity for establishing a measure of inter-rater reliability, this appears to be a reasonable instruction. The scale's purpose is to provide both researchers and educators with a means of measuring oral reading fluency with an emphasis on fluency in reading connected texts (Allington, 1983). Allington hoped to allow for greater precision in assessing fluency aside from a simple dichotomous distinction of reading being fluent or not. Thus, based on material published by Aulls (1978) and Allington (1983; Allington & Brown, 1979), it appears that the scale could be used either for informal assessment of students who struggle with fluent reading or more formal assessment, which would require multiple raters to ensure reliability.

Description of the scale's content. Aulls' (1978) original scale was a 7-point scale, but Allington's revision resulted in a scale with possible scores ranging from one to six on a single dimension. The scale is largely a continuum ranging from "word by word" reading (Allington, 1983, p. 559) to reading with appropriate phrasing, attention to punctuation, emphasis, and imitating normal speech. Aulls (1978) recommends regarding a score of one through four as not fluent while a score of five through seven indicates appropriate developing fluency. Based on the descriptions accompanying each score, Aulls' recommendations would correspond to Allington's scale as follows: a score of one through four indicates dysfluent reading, and a score of five or six indicates fluent reading. Because specific recommendations by Allington are not available, teachers and researchers who use this scale are likely to determine their own criteria based on curricular standards.

Instrument development. Aulls (1978) provides no information about the original scale's development. Allington and Brown (1979) modified this original scale for use in their reading program based on their understanding of fluency at the time. The common version of the scale (Allington, 1983) was adapted from the scale used in the reading program. It does not appear that any revisions were made by the author(s) after that point. Gathering psychometric information regarding the scale's development was attempted, but Richard Allington stated that because scientifically based reading research was not as common in the early 1980s, no psychometric data was available (personal communication, August 25, 2010).

While the scale's development relied mostly on the expertise and perception of the authors, this does not necessarily preclude it from being valid. Later studies using the scale provide more evidence of whether or not the scale is appropriate for measuring general fluency. In fact, Allington stated that because of their experience and expertise, teacher or researcher judgments of fluency may be just as adequate as ratings from a scale (1983). At the time, of course, there was very little research that provided recommendations about the details of fluent reading (cf. Clay & Imlach, 1971; Schreiber, 1980). It has been up to later studies to determine the validity and reliability of this rating scale.

Studies in which the scale has been used. Studies which have used Allington's fluency scale are rare; this is possibly because studies need to provide validity and reliability evidence for the measures used, and this information is not readily available. Thus, researchers may be unwilling to risk the possibility of using an unreliable and/or invalid instrument. Because the scale, though, has been used in a few studies (Rasinski, 1985; Young & Bowers, 1995; Young et al., 1996), it is possible to gauge the general potential of the scale for further use.

Rasinski (1985) used a slightly modified version of Allington's scale—he did not change the point values or the scaling, but simply added two more dimensions to the rubric. Thus, the scale effectively remained in its basic form. Rasinski assessed multiple reading sub-skills of third and fifth grade students, and he also assessed their general comprehension. The purpose of his study was to evaluate multiple variables and their contribution to reading fluency; to determine the relationships among these variables during reading; and to test models describing the relationship between automatic word identification, connected-text reading, and phrasing in contributing to reading comprehension. Rasinski used scores from Allington's scale as the measure for his phrasing variable.

Young and Bowers (1995) gave poor and average fifth grade readers easy and difficult texts to read aloud. They used Allington's (1983) scale to measure overall reading fluency and examined the relationship between students' word identification skills and text phrasing (among other variables) on their oral reading fluency and expressiveness. Of particular interest, they examined whether students' parsing ability—which they termed phrasal knowledge—played a role in fluency and expressiveness beyond what reading rate and accuracy could account for. In a following study (Young et al., 1996) authors examined the effects of repeated reading practice as well as prosodic modeling on fifth grade students' reading improvement. They used Allington's (1983) scale as a measure of reading fluency in addition to other measures of reading rate and accuracy.

A scale somewhat similar to the Allington (1983) scale was used by Tindal and Marston (1996) to measure students' reading expression. Teachers used the scale to rate students' reading expression in comparison to their peers with a score of *one* indicating that a student read with poor expression, *seven* indicating that a student read with excellent expression, and *four* indicating a typical performance for a student at a particular grade level within the school district. Students' scores were summative across four passages. Likewise, in a second study (Tindal & Marston, 1996) which tracked the progress of an individual student, a slightly modified version of the Allington (1983) scale was used.

Reliability evidence. While ideally multiple evidences of reliability will be gathered for any instrument (Crocker & Algina, 1986), because a student's score on a rating scale is largely determined by the rater, it is important to establish some level of consistency in ratings across raters if the score is to be deemed reliable (see Frick & Semmel, 1978). The goal is to get a good estimate of the student's true score within an acceptable error range. When students are rated based on their performance, it is important to know how well the rating given by one rater will generalize across multiple raters. Inter-rater reliability—typically measured as a correlation among two or more raters—is the most important type of reliability evidence that can be gathered for this type of instrument. Of course, it is important to note that scales which contain few points are likely to have higher inter-rater agreement than scales with more points. While there are methods available to account for these differences between scales, none of the studies described in this paper utilized such methods.

As no psychometric information is available from the development of either the original Aulls (1978) or more current Allington (1983) scale, reliability information had to be gathered from the two studies which have used this scale and reported reliability coefficients (Rasinski, 1985; Young et al., 1996) as well as another study which used a similar self-created scale in addition to a modified version of the Allington scale (Tindal & Marston, 1996). Unfortunately, Young and Bowers (1995) did not report inter-rater reliability, though they did mention that they used two raters for their study. Likewise, Tindal and Marston (1996) used only one rater to rate each student and gave no information regarding students' performance on individual passages, so neither inter-rater agreement nor test-retest reliability information was able to be gathered. In their follow up study which utilized a 5-point modified version of the Allington scale, Tindal and Marston tracked the progress of a single student but, again, reliability evidence was not gathered.

Rasinski (1985), on the other hand, reported two reliability indices: test-retest and inter-rater reliability. Rasinski reported a test-retest reliability coefficient of .90, but he did not provide any further details or specify how much time had elapsed between sessions. Nonetheless, a correlation of .90 is usually deemed adequate or high (Crocker & Algina, 1986). He reported inter-rater reliabilities of .96 for third grade readers and .98 for fifth grade readers with trained raters (the author does not elaborate on how much or what type of training took place). Rasinski did allow raters to confer with one another about readers' particular performance in relation to the rating criteria before actually rating students independently. This conferring is likely to have strengthened the agreement level among raters, but these numbers are, nonetheless, quite strong (Crocker & Algina, 1986). Perhaps with more extensive training, raters may be able to have high levels of agreement even without conferring with one another.

Young et al. (1996) also found high inter-rater agreement (94.4%) and reliability (r = .85) among two trained raters. The authors had raters independently rate 50-word sections of the 150-word texts they used in their study. The total rating for a text was then determined by averaging the three section ratings. Because the rating scale is an ordinal scale, if one rater's total text rating was 4, and the other rater's rating was 4.3, this would be counted as agreement. Ratings that varied by less than one point were counted as agreement. This could be problematic in cases where rounding would result in different ratings (e.g., a rating of 4.0 versus a rating of 4.8) even though the ratings differ by less than a point, but it is not wholly uncommon to count ratings that differ only by ±1 point as agreement (Crocker & Algina, 1986).

While inter-rater reliability is important to establish, it is not sufficient for estimating the reliability of an instrument. Inter-rater reliability simply indicates consistency among raters across a fairly diverse sample of participants and does not indicate the reliability of the instrument itself (Frick & Semmel, 1978). Thus, ideally, researchers who use rating scales such as the one developed by Allington (1983), would also obtain test-retest coefficients or a version of parallel forms reliability in which two or more ratings of the same student reading equivalent texts would be compared. This is difficult, however, due to the controversy surrounding the equivalency of texts in reading assessment (but see Christ & Ardoin, 2009). The reliability of an instrument is necessary for establishing the general validity of inferences made from the instrument, but reliability itself is not sufficient (Crocker & Algina, 1986; Frick & Semmel, 1978). Additional validity evidence must be gathered.

Validity evidence. According to classical test theory, there are three basic types of validity evidence that can be gathered for an assessment: content, criterion, and construct (Crocker & Algina, 1986; cf. Kane, 1992; Messick, 1995). Typically, studies use prosody or fluency scales to predict current and future reading comprehension—arguably falling under the domain of predictive-criterion validity. However, because theoretical models of reading tend to depict fluency as a construct that is strongly related to reading comprehension, one could argue that any predictive validity evidence may also fall within the category of construct validity—for if a scale that purports to measure prosody (an integral component of fluency) or fluency does not predict comprehension, then it is likely that the scale does not adequately reflect the construct (see Messick, 1995, for a discussion of construct validity).

Studies which have used versions of Allington's fluency scale (1983), have reported concurrent relationships with comprehension, reading rate, accuracy, and other measures as well (Rasinski, 1985; Tindal & Marston, 1996; Young & Bowers, 1995; Young et al., 1996). Rasinski (1985) found that fluency ratings of third graders correlated highly with reading rate (*r* = .862), accuracy (*r* = .649), and general comprehension (*r* = .738); ratings correlated moderately with retelling (*r* = .364) and multiple choice questions based on the passage read (*r* = .481). Correlations among fluency ratings and other reading skills for fifth grade students were similar: reading rate (*r* = .801), accuracy (*r* = .451), general comprehension (*r* = .729), retelling (*r* = .328), and multiple choice questions based on the passage read (*r* = .503). These relationships among variables indicate that scores on this scale, which focuses on the prosodic elements of reading aloud, relate highly with other indicators of reading fluency (i.e., reading rate and accuracy) as well as a standardized measure of general reading comprehension—the reading outcome measure.

Likewise, Tindal and Marston (1996) found that out of up to seven different reading measures, holistic fluency—as assessed by the Allington scale (1983)—and reading rate maintained the strongest consistent correlation with other reading skill indicators across grades one through six—including measures of comprehension, letter identification, and teacher judgments of reading skill. Young and colleagues (Young & Bowers, 1995; Young et al., 1996) also found strong relationships between scores on the Allington scale and other reading measures. Young and Bowers (1995) divided their sample into two groups and found—as expected—that students who scored higher on a general reading comprehension test also received consistently higher scores on the Allington scale across texts of varying difficulty levels. While poor readers' fluency degraded with increasing text difficulty, average fifth grade readers' fluency did not change across texts ranging from second to fifth grade reading levels. Based on theories of automaticity in reading fluency (e.g., LaBerge & Samuels, 1974; Perfetti, 1985), this is the expected outcome. Had more difficult texts been tested, then it would be expected that even the average readers' reading rate would begin to decline, though the effect of difficult texts on prosody, specifically, has not always had the same result (e.g., Benjamin & Schwanenflugel, 2010). Finally, Young et al. (1996) found that at different time points during an intervention study, fluency ratings correlated dependably with reading comprehension (ranging from *r* = .41-.52). Fluency ratings also correlated highly with words per minute (WPM; *r* = .82-.89) and moderately with accuracy (*r* = .52-.58, though the relationship was non-significant at one of the three time points).

While these studies illustrate the relationship that the Allington fluency scale ratings have with other related reading measures—important evidences of construct validity—it is now possible to directly measure many of the features that are described in Allington's scale (Benjamin & Schwanenflugel, 2010; Clay & Imlach, 1971; Cowie et al., 2002; Miller & Schwanenflugel, 2006; 2008). These studies which use direct measuring techniques have found that prosody has a moderately strong relationship with comprehension and a strong relationship with reading rate. Thus, the validity of scales like Allington's can and should continue to be assessed based on their inclusion of prosodic features that have been found indicative of good fluency and comprehension.

Summary evaluation. While the reliability of an instrument does not ensure the validity of inferences made from the results, reliability is a critical prerequisite for validity. Further research that demonstrates the stability of these rating measures over time and across equivalent texts should be conducted. Until further research is conducted, the scale likely remains generally valid for informal use and low-stakes assessment within classrooms. Its ease of use, little required training, and quick administration lend it to use in these informal settings. However, it lacks more precise psychometric review needed for more formal or high stakes use in both research and practice.

The NAEP fluency scale. The NAEP fluency scale was developed for the 1992 NAEP study of children's oral reading (Pinnell et al., 1995), sponsored by the National Center for Education Statistics. It was also used for the subsequent 2002 study (Daane et al., 2005) and has not been modified by either group of authors since its inception. Because these studies focused exclusively on fourth grade readers, the initial use of the scale was fairly limited. No specific age or grade-level requirement, however, is given by the authors, and the scale has been used widely among subsequent researchers and educators

Purpose of test and suggested uses. The scale should be used to assess the holistic fluency of the reader. That is, the scale takes into account the ease with which a student reads aloud—the phrasing, expression, and overall flow of the reading. The scale is not intended as a measure of accuracy or a precise assessment of reading rate. Rather, the focus is on assessing the "naturalness" and "effortlessness" that are characteristic of fluent reading (Pinnell et al., 1995, p. 14).

Description of test content. As a rating scale—or rubric—the NAEP Fluency Scale is descriptively simple. It consists of four levels to which raters can rate students' oral reading (see Table 2.2). Level 1 indicates that the student "reads primarily word-by-word. Occasionally two-word or three-word phrases may occur, but these are infrequent and/or they do not preserve meaningful syntax" (Pinnell et al., 1995, p. 15). In contrast, level 4 indicates that the student "reads primarily in larger, meaningful phrase groups. Although some regressions, repetitions, and deviations from text may be present, these do not appear to detract from the overall structure of the story. Preservation of the author's syntax is consistent. Some or most of the story is read with expressive interpretation" (p. 15). A score of 1 or 2 on the scale indicates that a student is not reading fluently; a score of 3 or 4 on the scale indicates that a student is reading fluently (Pinnell et al., 1995; Daane et al., 2005).

Instrument development. The NAEP Fluency Scale resulted from a theoretical perspective that emphasizes fluent reading as a process in which meaning is being constructed (Pinnell et al., 1995). The authors emphasized that in order to read fluently, students have to recognize more than just words. They have to be able to recognize and understand structure in larger units such as phrases. Only then can comprehension take place. Thus, students who understand what they are reading will likely reflect the structural organization intended by the author; additionally, their reading will reflect an understanding of larger story elements and concepts. All of these characteristics will work together to produce oral reading that is smooth, effortless, and has a sense of ease. Accuracy in reading was intentionally kept out of the scale, as the authors operated under

Table 2.2

NAEP Fluency Scale

Fluent	Level 4	Reads primarily in larger, meaningful phrase groups. Although some regression, repetitions and deviations from text may be present, these do not appear to detract from the overall structure of the story. Preservation of the author's syntax is consistent. Some or most of the story is read with expressive interpretation.
	Level 3	Reads primarily in three or four word phrase groupings. Some word-by-word reading may be present. However, the majority of phrasing seems appropriate and preserves the syntax of the author. Little or no expressive interpretation is present.
Nonfluent	Level 2	Reads primarily in two word phrases with some three or four word groupings. Some word-by-word reading may be present. Word groupings may seem awkward or unrelated to larger context of sentence or passage.
	Level 1	Reads primarily word-by-word. Occasionally two or three word phrases may occur, but these are infrequent and/or they do not preserve meaningful syntax.

Note: From the U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000, Oral Reading Study cited in Daane, M.C., Campbell, J.R., Grigg, W.S., Goodman, M.J., & Oranje, A. (2005).

the belief that accuracy is separate from reading fluency (Pinnell et al., 1995; Daane et al., 2005).

Based, then, on their theoretical rationale and a review of the literature available at the time, the authors developed a 4-point scale varying from "word by word" (Pinnell et al., 1995, p. 15) reading to reading in appropriate units and reflecting the intentions of the author. The increase of point values on the scale reflects the authors' belief that varying degrees of comprehension will be demonstrated by particular qualities in a student's oral reading. Thus, a student who reads at level 3 is likely comprehending significantly more than a student who reads at level 1. Once this scale was developed, there is no indication that any revision took place. In fact, the exact same scale was used in the later 2002 NAEP study (Daane et al., 2005).

Studies in which the scale has been used. The NAEP Fluency Scale has proven to be largely popular as a classroom tool for informal reading assessment (McKenna & Stahl, 2009). It has also, though, been used in several large and mid-sized research studies (Daane et al., 2005; Kuhn, 2005; Pinnell et al., 1995; Valencia et al., 2010). While the NAEP studies exclusively tested fourth grade students, the scale certainly has not been limited to use with that age group. McKenna and Stahl (2009) include this scale in their book which provides reading assessments and recommendations that can be applied to students of all ages who are learning to read or struggling with reading. It seems, then, that this scale—like Allington's (1983)—is appropriate for broad application.

The large NAEP studies (Daane et al., 2005; Pinnell et al., 1995) used the scale as an overall indicator of the reading fluency of well over 1,000 fourth graders serving as a national sample. The authors specifically stated that accuracy was not taken into account in the scale. Separate measures of fourth graders' reading rate and accuracy were also taken in addition to numerous other factors. Thus, in these studies reading fluency was seen as a separate construct from reading rate and accuracy (cf. Kuhn et al., 2010). The purpose of these studies was to review children's progress in reading at a national level, and then provide wide-scale recommendations for improvement.

Kuhn (2005) compared two popular methods of group fluency practice in second grade classrooms: repeated reading and wide reading. Kuhn selected six struggling readers for each of four intervention groups: a control group, a listening-only group (control for the Matthew Effect), a wide-reading group, and a repeated reading group.

Among a battery of assessments measuring word-reading, reading rate, accuracy, and comprehension, Kuhn used the NAEP Fluency Scale as a measure of prosody at the beginning and end of the intervention. Thus, the assumption was that the scale was sensitive enough to detect changes in prosody over a six week intervention period.

Finally, Valencia et al. (2010) examined the validity of using WCPM as a measure of oral reading fluency and as a tool used in educational decision making. Using second, fourth, and sixth grade children, the authors compared the performance of WCPM to other individual and combined indicators of passage comprehension and oral reading fluency—such as rate, accuracy, and prosody ratings—in predicting general reading comprehension. Likewise, authors compared the performance of WCPM to other fluency indicators in identifying students at risk for reading failure. Valencia et al. used the NAEP Fluency Scale as the prosody measure across the three grade levels, but since the scale was originally designed with fourth grade students in mind, they adapted the scale for each of the three grade levels they studied. Valencia et al. did not report their methods used in adapting the scale and did not report the actual adapted scales.

Reliability evidence. The criteria for establishing reliability for the NAEP Oral Reading Fluency Scale are identical to those described above under the Allington (1983; Allington & Brown, 1979) scale. Inter-rater reliability is often the only type of reliability information provided in studies using fluency rating scales, though test-retest and alternate forms reliability coefficients are desirable as well (Crocker & Algina, 1986). However, because these scales are often used in classrooms by various teachers and reading specialists, it is certainly critical to establish an acceptable level of inter-rater reliability before using the scale widely.

Inter-rater reliability for the NAEP scale has increased since it was first developed. In the original study using the scale (Pinnell et al., 1995), a reliability coefficient (the type of coefficient was not reported) of .70 was reported with only 58% exact inter-rater agreement, but 98% adjacent agreement. Since the scale is only composed of four levels, exact agreement does not seem an unreasonable expectation, but agreement and reliability reported in this study are below desired levels (Crocker & Algina, 1986; Frick & Semmel, 1978).

Higher agreement and reliability levels were achieved in later studies. The second NAEP oral reading study reported an intraclass correlation coefficient of .82 as well as 81% exact inter-rater agreement and 100% adjacent agreement (Daane et al., 2005). Since a score of 1 or 2 on the scale indicates dysfluent reading and a score of 3 or 4 indicates fluent reading, researchers also looked at whether raters agreed in their classification of children as fluent or dysfluent. The 1992 study (Pinnell et al., 1995) reported 74% classification agreement, while the 2002 study (Daane et al., 2005) reported 92% classification agreement. It appears that similar rater training methods were used in both studies. Perhaps within the 10 year space between studies, teachers and reading specialists in general became more familiar with the NAEP scale and used it in their schools.

The much smaller study conducted by Kuhn (2005) relied on the NAEP scale as a measure of prosody in children's reading. While 20 raters were used for the 2002 NAEP study (Daane et al., 2005), only two expert raters were used in the Kuhn (2005) study—the researcher and a colleague (Melanie Kuhn, personal communication, November 5, 2010). 100% inter-rater agreement was reported, though rater training was not discussed and a reliability coefficient was not reported. However, 100% agreement in a small study in which a reliability coefficient may have been affected by range restriction or other factors is certainly greater than adequate.

Valencia et al. (2010) trained their raters using selected audio samples of children's reading that spanned a wide range of abilities for each of the three grade levels (2nd, 4th, & 6th) that they examined. Raters listened to the first three minutes of reading for each child. Like in the previous NAEP studies, readings with scores of 1 or 2 were deemed not fluent while readings with scores of 3 or 4 were deemed fluent. Unlike any of the previous studies using the scale, raters assigned a scale score to each line of text read by the student, and then based the overall score on the line scores. Researchers reported 82% inter-rater reliability, though because it is reported as a percentage it is not clear whether this 82% refers to inter-rater agreement or a reliability coefficient. Regardless, either as agreement or reliability, this number is similar to that reported by Daane et al. (2005).

Because the scale is restricted to only four levels, it seems that traditional agreement percentages should be quite high, as in Kuhn (2005). However, the agreement levels reported in these studies using the NAEP scale are lower, on average, than those reported in studies using the Allington scale (Rasinski, 1985; Young et al., 1996). This may be due to the lengthy descriptions at each level on the NAEP scale and the fact that most characteristics described in the scale (e.g., reading in two to three word phrases) overlap across levels, simply varying by degree. While the Allington scale uses this type of overlap as well, the Allington scale has more levels and levels appear to be more segregated from each other. For example, in examining the NAEP scale, it may be somewhat difficult to distinguish between level 1 and level 2 as both describe poor oral reading. A level 1 reader may sometimes read in phrases whereas a level 2 reader simply reads in phrases a little more frequently, and both levels indicate that meaningful structure is not being preserved. The apparent lack of distinctiveness at these levels may result in disagreement among raters, and except for the expert ratings conducted in the Kuhn (2005) study, it appears that inter-rater reliability has consistently been less than optimum.

Validity evidence. As with the Allington (1983; Allington & Brown, 1979) scale, most studies using some form of the NAEP scale have used it to predict reading comprehension and have provided evidence of the fluency ratings' relationships with other indicators of reading skill. As would be expected in a large study of children's reading abilities, ratings from the 1992 NAEP oral reading fluency study (Pinnell et al., 1995) approximated a Normal curve, with small percentages of students earning ratings of 1 or 4 (7% and 13%, respectively), and the majority of students earning ratings of 2 or 3 (37% and 42%, respectively). Results were fairly similar in the later NAEP oral reading study using a different text (Daane et al, 2005): relatively few students earned ratings of 1 or 4 (8% and 10%, respectively) while greater numbers of students earned ratings of 2 or 3 (32% and 51%, respectively). However, the scale was not designed to be a normative assessment; rather, if children were reading fluently, they would get scores of 3 or 4. Researchers found that even the generally fluent fourth grade students rarely read with the level of expressiveness required for a level 4 rating (Daane et al., 2002; Pinnell et al., 1995). This presents a potential weakness in a more criterion-referenced assessment by setting a standard that may not be necessary or even possible for most students to reach.

The fluency scores did have strong relationships with students' other reading performance and experiences (Pinnell et al., 1995). Children who reported reading more books outside of school tended to have higher fluency scores, and fluency scores were positively related to reading accuracy and very strongly related to reading rate. Additionally, fluency scale scores were positively related to students' average reading proficiencies as determined from the larger NAEP reading studies (Daane et al., 2005; Grigg, Daane, Jin, & Campbell, 2003; Mullis, Campbell, & Farstrup, 1993; Pinnell et al., 1995). These relationships support the validity of inferences made using the NAEP scale, though some modifications may be necessary once research determines just how expressive children can and should be in their oral reading. Because of the limited scope of the scale (it was designed to determine whether a child is fluent or nonfluent versus functioning more as a diagnostic tool for reading instruction) only basic inferences regarding a student's current level of oral reading fluency can be supported.

Kuhn (2005) did not have enough second grade participants to run inferential statistical analyses, but did find that students in her intervention study who received explicit fluency instruction through repeated reading and wide reading methods improved in their NAEP scores over a six-week period more than children who only received their typical instruction. This finding supports the claim that the NAEP scale is actually measuring what it proposes to measure—holistic fluency. Valencia et al. (2010) conducted many analyses and found, as expected, that fluency ratings increased as a function of increased grade level (grades 2, 4, & 6). They also found that these ratings were highly related to WCPM (*r* = .84 for each grade level) and also served as moderate predictors of comprehension in several structural equation models—though the usefulness of these models are dubious as no indices of model fit were reported.

It is difficult to determine how useful the NAEP scale is when only a few studies have reported statistics related to its construct and predictive validity. However, because of the large-scale nature of the NAEP studies (Daane et al., 2005; Pinnell et al., 1995), I am fairly confident that the scale can be used as a general indicator of whether a child is fluent or not in his or her oral reading. It is positively related to other indicators of reading fluency like accuracy and reading rate, and it has been shown to predict overall reading ability (Daane et al., 2005; Pinnell et al., 1995), which includes comprehension. However, the scale is too broad in description and compact in design to serve as an effective instructional tool for addressing specific student needs in oral reading skill.

Summary evaluation. The NAEP scale is at first glance easy to use and only requires moderate training to implement. It provides teachers and researchers with a quick means of assessing oral reading fluency largely in terms of expressiveness and phrasing. Recent studies have found fair to good inter-rater reliability, though other types of reliability (such as within-rater [test-retest] reliability or alternate forms reliability) have not been tested. Additionally, in its basic form, I was only able to find three studies using the scale with fourth graders and one study with second grade students. Valencia et al. (2010) modified the scale for their second grade and sixth grade participants, though other research has not yet determined if such modifications are necessary. The scale appears to be valid for making basic determinations about whether a child's oral reading if fluent or not. However, further inferences—such as what type of instruction may be needed—based on scale scores seem impossible due to the broad-strokes nature of the scale. In addition, for the scale to be considered a true holistic measure of fluency it would need to better match current definitions of oral reading fluency as a multidimensional construct, including reading rate, accuracy, and prosody.

The Multidimensional Fluency Scale. The Multidimensional Fluency Scale was originally developed and published by Zutell and Rasinski (1991), though it got its start when Rasinski slightly modified Allington's (1983; Allington & Brown, 1979) scale by adding several additional descriptive dimensions to it (without changing point values) for his dissertation research (Rasinski, 1985). Through the years, several permutations have arisen: the 1991 version of the scale allowed educators to rate readers on three dimensions of fluency: phrasing, smoothness, and pace. In 2004 an adaptation was published (Rasinski, 2004) which was based on four dimensions of fluency: expression and volume, phrasing, smoothness, and pace. Finally, Rasinski and his colleagues more recently published a modified edition of the scale (Rasinski et al., 2009), termed the "Multi-Dimensional Fluency Scoring Guide." This scale, like the 1991 version, assessed readers on three dimensions: phrasing and expression, accuracy and smoothness, and pacing. This accuracy component is the newest addition to the scale, and appears to fall in line with many definitions of fluency which incorporate accuracy, reading rate, and prosody (e.g., Hudson et al., 2005; Kuhn et al., 2010; NICHHD, 2000). Including accuracy in the scale also sets the MDFS apart from scales like the NAEP and Allington scales, discussed above.

There is no evidence that the scale is intended for a particular limited group of student readers, as Rasinski himself used the scale (or a similar one) for students ranging from third grade through middle school (Rasinski, 1985; Rasinski et al., 2009). So it appears that the scale is designed to be used with a wide range of children, those who have acquired enough basic decoding and word recognition skills to begin working on fluent reading versus simply word identification.

Purpose of test and suggested uses. The purpose of the scale, as described in Rasinski (2004), is to assess reading fluency. The author especially focuses on the scale as a method for assessing prosody or expression—something which, unlike reading rate and accuracy, is not easily measured using purely quantitative means. Thus, the scale fills a void that is left when only rate and accuracy measures are used to assess a child's fluency. The author claims that a 60 second reading sample (or less) is all that is necessary to make a reliable and valid assessment of the child's fluency. While Zutell and Rasinski (1991) instruct teachers to use instructional-level materials when teaching fluency to their students, Rasinski recommends using the scale to assess a child's oral reading fluency with a grade-level passage (2004). Thus, to effectively rate students using the MDFS, teachers should listen to a brief sample of the child reading a grade level text.

Description of test content. The most recent iteration of the MDFS (Rasinski et al., 2009) is comprised of three dimensions (see Table 2.3), with four levels within each dimension. The final rating, then, is the sum of the ratings of levels across the three dimensions with a possible total score ranging from 3-12. Previous versions of the scale differ somewhat in their descriptions of the dimensions and levels. The original MDFS (Zutell & Rasinski, 1991) is quite similar to the current version and is also comprised of three dimensions, while a later adaptation (Rasinski, 2004) includes four dimensions rather than three. For this four dimensional version, the author states that a score lower

Table 1.3

2009 Multidimensional Fluency Rubric

A. Phrasing and Expression	
4	Reads with good expression and enthusiasm throughout the text. Sounds like natural language throughout the text. Varies expression and volume to match his or her interpretation. Generally well-phrased and meaningful; mostly in phrase, clause, and sentence units, with adequate attention to expression.
3	Makes text sound like natural language throughout the better part of the passage. Mixture of run-ons, mid-sentence pauses for breath, and possibly some choppiness; reasonable stress/intonation.
2	Begins to use voice to sound like natural language in some areas of the text but not others. Generally two and three word phrases, which break up the reading; improper or inadequate stress and intonation; fails to mark tile ends of sentences, clauses, and phrases.
1	Little sense of trying to make text sound like natural language. Tends to read in a quiet voice and/or monotonic, unenthusiastic reading, with little sense of phrase boundaries. May be frequent word-by-word reading or run-on word calling with no attention to expression.
B. Accuracy and Smoothness	
4	Generally smooth and accurate reading with a few decoding breaks; word and structure difficulties are resolved quickly, usually through self-correction. Smooth phrasing enhances the interpretation.
3	Occasional decoding breaks in smoothness caused by difficulties with specific words and/or syntactic structures. Additions or deletions are minimal and usually resolved. Smoothness includes attention to phrases.
2	"Rough spots" in text where extended pauses, hesitations, sound outs, etc., may be frequent and disruptive. Student may add or delete words without correcting. There may be a combination of rough and smooth spots with little attention to phrasing.
1	Extended pauses, hesitations, sound-outs, repetitions, or multiple attempts MAY be present. Words may be changed, added and/or deleted without notice. The reader does not attend to the smooth delivery of phrases.
C. Pacing	
4	Consistently conversational; appropriate rate throughout reading. Pace enhances the meaning of the text.
3	Mixture of appropriately quick and overly slow or fast reading. Attention to the effect of pace on meaning throughout most of the text.
2	Generally inappropriate speed for the text. Pays little attention to the effect of speed on the meaning of the text.
1	Inappropriate speed; slow and laborious or run-on. Ignores the effect of speed on the meaning of the text.

Note: From 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.

than 8 is indicative of poor oral reading fluency while scores of 8 or above are indicative of adequate progress in oral reading fluency.

Instrument development. Originally, the scale was developed simply by modifying other existing scales (Allington & Brown, 1979; Aulls, 1978; Rasinksi, 1985) based on the current literature. Rasinski and colleagues continued to modify the MDFS based on feedback from educators as well as observations of reader behaviors (Zutell & Rasinski, 1991). The first true MDFS was published in 1991 (Zutell & Rasinski, 1991) and the rationale given for creating multiple dimensions is that the authors observed that readers sometimes performed adequately on one aspect of fluent reading (e.g., phrasing) but poorly on another (e.g., smoothness). Additionally, they found that when teachers used single-dimension rating scales, some teachers weighted certain features more highly than others, and agreement difficulties tended to lie in the midpoints of the scales.

While the 1991 version of the MDFS was comprised of three dimensions (phrasing, smoothness, and pace) a 2004 edition included an additional dimension: expression and volume. No rationale for this modification was given, though the author may have been inspired by the recently-created NAEP fluency scale (Pinnell et al., 1995), which included expression as a factor in fluent reading. In a later adaptation of the scale (Rasinski et al., 2009), authors describe it as a “rubric” (p. 355) and describe it as an elaboration of the NAEP scale (Daane et al., 2005; Pinnell et al., 1995). The NAEP scale, then appears to have had significant influence in this latest iteration of the MDFS.

The development of the original MDFS (Zutell & Rasinski, 1991)—whose genesis really began with Rasinski (1985)—was the result of careful observation of readers and feedback from teachers in addition to an understanding of the fluency literature. The scale, then, was a step forward in fluency assessment. However, later revisions of the scale are more difficult to understand based on published descriptions. No explanation is given for the modified 2004 version (Rasinski, 2004), and Rasinski et al. (2009) state that the most recent version was inspired by the NAEP scale with no mention of the earlier Aulls (1978) or Allington (1983; Allington & Brown, 1979) scales. Additionally, no explanation is given for the dimensional changes in the scale over time: the 1991 version contained three dimensions; the 2004 version contained four dimensions; and the 2009 version contained three dimensions. While the modification process is a bit confusing, overall, the scale is easily identifiable throughout its iterations and maintains its multidimensional status.

Studies in which the scale has been used. Of the existing fluency scales discussed in this article, the MDFS appears to have been used in more research studies than the others. However, its popularity among teachers compared to the NAEP and Allington scales is unknown. One study was found that used the 1991 version of the MDFS (Sargent, 2002); several studies used the 2004 version of the scale (Clark, Morrison, & Wilcox, 2009; Reutzel, Fawson, & Smith, 2008; Yılıdz, Yıldırım, Ateş, & Çetinkaya, 2009); and one study has used the most recent version of the scale (Rasinski et al., 2009).

Sargent (2002) conducted a study with fifth graders, examining various measures of reading skill and oral reading fluency to find which measures best predicted current and future scores on a standardized criterion-referenced test of reading. One of the oral reading fluency measures used was the 1991 iteration of the MDFS. However, only 52 students participated in the study, and all students came from the same school.

Several studies utilized the four-dimension 2004 version of the scale. Reutzel et al. (2008) wanted to determine whether scaffolded silent reading worked as well as guided repeated oral reading with feedback in promoting third graders' development of reading fluency. They looked for reductions in error rates, increases in WCPM, increases in expression ratings (as measured by the MDFS) and increases in passage recall as evidence of improvement. Yılıdz et al. (2009) examined the fluency skills of fourth grade Turkish children, focusing specifically on comparing students' WCPM performance with their prosodic reading performance as measured by the MDFS. Finally, Clark et al. (2009) attempted to isolate particular aspects of oral reading fluency that developed among different students involved in a reader's theater intervention. This study was rather small (using only three fourth graders) but students were selected based on a differentiation of their reading abilities and authors provided detailed analyses of students' progress.

The most recent 2009 edition of the MDFS was developed by the Educational Service Unit #3 (ESU #3, an education agency in Omaha, Nebraska) and used by Rasinski et al. (2009) in the paper which introduced this version of the scale. Rasinski and colleagues attempted to demonstrate that fluency is a reading skill that should not simply be relegated to the primary grades and even in older children should not simply be measured by reading rate. The authors proposed that the relationship between fluency and comprehension should remain significant even as grade level increases, though the relationship will diminish some. Approximately 400 children in each of three grades (3rd, 5th, and 7th) participated in this study.

Reliability evidence. The criteria for establishing reliability for the MDFS are identical to those described above under the Allington (1983; Allington & Brown, 1979) scale. Inter-rater reliability is often the only type of reliability information provided in studies using fluency rating scales (though sometimes only inter-rater agreement percentages are provided) though within rater test-retest and even alternate forms reliability coefficients may be desirable as well (Crocker & Algina, 1986). However, because these scales are often used in classrooms by various teachers and reading specialists, it is critical to establish an acceptable level of inter-rater reliability before using the scale widely.

Reliability evidence was not reported for either the 1991 (Zutell & Rasinski, 1991) nor the 2004 (Rasinski, 2004) versions of the scale, and outside studies which used the scale did not report reliability indices. Sargent (2002) used the 1991 version of the scale in his study of fluency and fifth graders, but no reliability data was reported. Using the four-dimensional 2004 edition of the scale, Reutzel et al. (2008) examined the effects of scaffolded silent reading versus guided repeated oral reading in developing third graders' oral reading fluency. They cited Zutell and Rasinski (1991) in claiming that the scale had a .99 inter-rater reliability coefficient, but in fact when Zutell and Rasinski (1991) discuss a .99 inter-rater reliability coefficient they are referring to a much earlier version of the scale from Rasinski (1985), which was actually a simple modification of the Allington scale (Allington, 1983; Allington & Brown, 1979). Reutzel et al. (2008) did not provide any reliability information from their own data. They had four raters rate each oral reading and took an average of the four ratings as the final score for each child rather than conducting a reliability study among raters. Yılıdz et al. (2009) compared

WCPM in fourth grade Turkish children with scores from the 2004 MDFS, but did not report inter-rater or any other type of reliability evidence. Clark et al. (2009) conducted an in-depth study of three fourth graders using a case-study design. They used the 2004 version of the MDFS, but no reliability data was reported.

When the latest version of the scale was presented by Rasinski et al. (2009) in the context of a large fluency study, they used several means to ensure reliability of data gained using the scale. Teachers and reading specialists were trained on each of the scale's three dimensions, and practiced by working in small groups until agreement was achieved. Additionally, the scale was tested by having two raters rate 6,000 elementary level reading samples; they obtained exact or adjacent inter-rater agreement of 94%. For the actual study conducted by Rasinski et al. (2009), two raters scored each rating sample independently. If the raters disagreed by more than one point on any of the three dimensions, the reading sample was sent to a third rater. This happened rarely. Inter-rater agreement was defined as plus or minus two on the 12-point scale and was reported as .857 for the study. It is unclear whether this number indicates a reliability coefficient or a simple agreement percentage, but it is likely that enough variance in scores exists so that agreement and reliability levels would be similar in such a large study.

It is unfortunate that so little evidence of reliability is reported in studies using the MDFS. Without such information, it is impossible to judge the appropriateness of the scale for use in classrooms or research. Also, there is little to guide researchers in determining the criteria for establishing reliability for this scale. The only study reporting reliability evidence (Rasinski et al., 2009) uses a criterion of plus or minus two points on the aggregate scale for agreement. This may be too lax for determining inter-rater agreement when reliability indices exist which take near-agreement into account. In short, outside researchers should be reporting reliability coefficients when using fluency rubrics in research studies. If research is to guide classroom practice, then research needs to provide educators with the tools necessary to effectively assess their students' skills.

Validity evidence. Only two the studies using the MDFS used it mainly as a means of predicting reading comprehension (Rasinski et al., 2009; Sargent, 2002). Rasinski et al. (2009) also had other experts review the scale for content validity. Other studies used the scale to demonstrate the effectiveness of a particular reading intervention (Clark et al., 2009; Reutzel et al., 2008) or to illustrate the connection that ratings from a fluency scale have with more traditional fluency assessments like WCPM (Yılıdz et al., 2009). Thus, issues of predictive-criterion and construct validity were explored in these studies.

Rasinski et al. (2009) found that reading fluency—as measured by the MDFS—shared significant variance with reading comprehension at all tested grade levels (r^2 = .402 for grade 3; r^2 = .432 for grade 5; r^2 = .326 for grade 7). Information regarding the normality of the fluency score data are not provided, but means and standard deviations are—based on a possible aggregate score ranging from 6-24, means and *SD* were 16.78 (4.62), 18.40 (4.22), and 17.79 (4.32) for grades 3, 5, and 7, respectively. Scores for all grades ranged from a low of 6 to a high of 24, and the standard deviations demonstrate that it is indeed possible for a large number of children to score highly on the rubric—compared to the very small number of fourth graders who obtained a score of 4 on the NAEP scale. This possibility of achieving scores in the highest ranges is important for rubrics and scales that are criterion-referenced. Finally, content validity was achieved by submitting the scale to a panel of five experts in reading. The panel was unanimous in its agreement that the scale was appropriate and valid for assessing oral reading prosody (Rasinski et al., 2009).

Sargent (2002) also examined the relationship between reading comprehension and scores on the MDFS in fifth graders, though in this case fall scores on the MDFS were used to predict spring comprehension scores. A different version of the scale was used (Zutell & Rasinski, 1991) as well as a different standardized comprehension test from that used in Rasinski et al. (2009), but a moderate correlation was still found between MDFS scores and general reading comprehension over time (r = .49; r^2 = .24). Unfortunately, the study was small (52 participants from a single school) so the findings are not as robust as those in Rasinski et al. (2009) or other large studies. However, these findings do support the predictive-criterion and construct validity of the MDFS.

If students are given specific instruction over time to improve their reading fluency using tested methods, one would expect that scores on a valid holistic fluency scale would reflect students' fluency improvements over time. Two studies were found which used the 2004 version of the MDFS as a means for measuring improvements in fluency over the course of an intervention (Clark et al., 2009; Reutzel et al., 2008). Reutzel et al. (2008) found that third graders significantly improved their MDFS scores over the course of a fluency intervention. Students also made similar gains in their reading rate and accuracy scores as well as their comprehension. Conducting a case-study examination of three fourth graders' progress, Clark et al. (2009) found that all three students showed marked improvement in specific fluency dimensions over the course of an eight week intervention. Additionally, these improvements—e.g., in expression—took place even if improvements in WCPM did not. This study demonstrates that while prosody, rate, and accuracy are all components of fluency, they do not measure exactly the same thing. Students can improve in one aspect but not another, and the fact that the MDFS was able to detect particular improvements in various aspects of fluency lends to arguments of its construct validity.

Yılıdz et al. (2009) used the 2004 version of the MDFS to compare Turkish fourth graders' holistic fluency performance with their WCPM. They claimed that while the children performed at grade level based on WCPM norms, they did not perform as well on the MDFS. Scores below 8 indicated that students were struggling with fluency while scores of 8 or above indicated that students were developing adequately. With a normal WCPM range of 70-110, students averaged 87.17; students' mean score on the MDFS was 8.97, considered adequate. However, the authors break the MDFS results down further by showing that 40% of students demonstrated problematic levels on the MDFS and thus claim that more targeted instruction is necessary. However, authors did not provide a breakdown analysis of the WCPM scores, so there is no evidence to support the claim that students performed less well on the MDFS. The similarity in overall performance of both fluency assessments (r = .74) attests to the construct validity of the MDFS.

While none of the above studies were specifically designed to test the validity of the MDFS, the overall evidence from the studies supports an argument for the validity of inferences made from MDFS scores. It is additionally interesting to note that the multi-dimensional structure of the scale seems to allow for more sensitive detection of specific fluency improvements during intervention (Clark et al., 2009)—unlike more general fluency scales such as the NAEP—though further studies would be needed to replicate these findings with a larger sample of participants. Overall, while the MDFS certainly requires more effort on the part of the rater, it may be the most informative of the three scales discussed in this paper.

Summary evaluation. In studies conducted by the author, teachers underwent extensive training to learn how to rate readers using the scale. Raters were trained on anchor reading samples and practiced rating in small groups, working until raters reached agreement (Rasinski et al., 2009). Further studies are necessary to determine how involved the training process would need to be to establish reliability among teachers within schools for classroom use. Additionally, while a simple 60 second reading sample may be all that is necessary to effectively rate readings using the scale, the fact that educators have to give ratings on three or four different dimensions of fluency will mean that the MDFS requires more time than either the Allington or NAEP scales. For simply a general look at children’s overall fluency, these simpler scales may be appropriate; however, if a more detailed examination of fluency strengths and weaknesses is desired, the MDFS will likely serve the purpose better. Unfortunately, there is simply not enough evidence, yet, to support claims that the scale can be used reliably among teachers; further research will need to bear this out.

Existing scales and research-based features of oral reading fluency. The scales discussed above vary in their focus on the multiple indicators of oral reading fluency, as defined by Kuhn et al. (2010). The Allington and NAEP scales focus largely on automaticity and prosody, while the Multidimensional Fluency Scale uses its multidimensional structure to focus on automaticity and prosody as well as accuracy.

Based, then, on a simple comparison of the scales with definitional components of reading fluency, the Multidimensional Fluency Scale appears to most closely represent fluency as a whole construct. That is, it is designed to stand alone as a fluency assessment.

However, none of the scales discussed above underwent extensive componential testing during their development. The NAEP scale is a holistic scale, and thus, is not divided into various components. The Allington scale was developed without psychometric testing. Finally, the Multidimensional Fluency Scale was initially published in 1991 (Zutell and Rasinski), and there is no indication in the literature that its prosodic features were tested individually for inclusion in the scale. Since these scales were constructed prior to the major expansion of research into oral reading prosody during the past decade, it is likely that a new scale—developed based on current knowledge of the literature and tested rigorously for its validity—will better serve the needs of modern educators and researchers.

Developing an Oral Reading Fluency Rating Scale

Methods for assessment development. Numerous texts and papers discuss the topic of test construction in general, but purposes of an assessment tend to dictate the actual steps taken in development. Crocker and Algina (1986) list basic principles for test construction in accordance with classical test theory, and many of these principles have carried over in some form to recommendations for creating classroom performance assessments (e.g., Gronlund & Waugh, 2009; Miller, Linn, & Gronlund, 2009; Nitko, 2004; Reynolds, Livingston, & Willson, 2009). Because an oral reading fluency rating scale is likely to be used as both a classroom performance assessment as well as a tool used in research, principles of formal test construction as well as recommendations for developing classroom assessments should be considered.

Based on assessment development recommendations from multiple sources, once the topic or construct for assessment is determined, and the method of assessment (i.e., a rating scale) chosen, a new oral reading fluency rating scale could be created and tested using the following series of steps:

1. Determine the specific behaviors/skills that represent the construct.
2. Identify the proportional focus that should be placed on each type of behavior/skill.
3. Select the type of scale that is most appropriate for measuring these skills.
4. Construct an initial scale, using between 3 and 7 rating positions for each dimension of the scale.
5. Make sure points on the scale are clearly defined.
6. Have the scale reviewed (revise as necessary).
7. Test the scale on a representative sample.
8. Revise and test again as necessary.
9. Conduct reliability and validity studies on the scale.
10. Create guidelines for administering, rating, and interpreting results.

Validation framework. Inferences made from an assessment must be backed up by evidence demonstrating that the assessment is an appropriate tool for making such inferences. While both Messick (1989; 1995) and Kane (1992) among others describe thorough and often-used approaches to validation, Kane’s argument-based approach allows for flexibility in making validity arguments in multiple contexts (Kane, 2006) and is intuitively sound, allowing for the researcher to make hypotheses about how certain evidence supports the assessment's validity. Kane's framework stresses the importance of making inferences from test results based on evidence outlined in interpretive arguments. There are two major components to this approach: the interpretive argument and the validity argument (Kane, 2006). The interpretive argument focuses on inferences and assumptions leading to the statements and decisions that can be made from assessment results. The validity argument evaluates the interpretive argument as a whole, and then the inferences and assumptions in the interpretive argument specifically using appropriate evidence (Cronbach, 1988). While the arguments can never serve to "prove" any conclusions about the validity of inferences made from test results, it can and should be convincing and plausible (Kane, 1992; 2006). Kane outlines four steps necessary for the argument-based approach to validation based on the interpretive argument framework:

> One (a) decides on the statements and decisions to be based on the test scores, (b) specifies the inferences and assumptions leading from the test scores to these statements and decisions, (c) identifies potential competing interpretations, and (d) seeks evidence supporting the inferences and assumptions in the proposed interpretive argument and refuting potential counterarguments. (p. 527)

This argument based approach lines up with the most recent recommendations provided in *Standards for Educational and Psychological Measurement* (American Educational Research Association [AERA] et al., 1999).

Based on Kane's framework, the evidence gathered to make an argument regarding a test's validity should be based on the inferences and assumptions made in the argument for how test scores should be interpreted. In the case of an oral fluency rating scale, then, one can first decide what types of statements should be made about a student's fluency skill based on the score he or she receives, and what decisions, if any, should follow. Second, the connection between these interpretive statements and the scale scores must be explained. In the case of oral reading fluency, inferences and assumptions case be explained in light of theories of reading development, prior research in reading fluency, etc. Third, relevant potential competing interpretations of a student's score should be considered. For oral reading fluency tests, issues such as text difficulty, language issues, and physical disabilities, among others, should be taken into account. Finally, the most time and resource intensive step should be taken. To make the validity argument evidence should be gathered that supports the interpretive argument and weakens or refutes opposing arguments.

Purpose of the Present Study

The basic goal for the present research is to test the usefulness of a new assessment tool for measuring oral reading fluency as a whole construct. Oral reading fluency is a complex construct comprised of three basic components: reading rate, word reading accuracy, and prosody (Kuhn et al., 2010). Prosody, specifically, has been of interest in recent research as modern technology has made precise prosodic measurements possible. Also, though, prosody's role in reading development and skill is less understood than either rate or accuracy and numerous methods are still being used to measure prosodic skill in children (e.g., Benjamin & Schwanenflugel, 2010; Clay & Imlach, 1971; Cowie et al., 2002; Ravid & Mashraki, 2007; Whalley & Hansen, 2006; Zutell & Rasinski, 1991). Fluency is both theoretically (LaBerge & Samuels, 1974; Logan, 1997) and empirically linked to comprehension (Fuchs et al., 2001), and is often used to predict children’s reading comprehension when time is limited or valid and reliable comprehension measures are not readily available.

As a reading skill in its own right, though, oral reading fluency is critical especially in reading development during the early elementary school years (Chall, 1983). Due to fluency’s critical role in the development of reading ability and its usefulness as a method for monitoring children’s overall reading skill, numerous assessments of fluency have been developed to measure components of fluency like reading rate and accuracy (e.g., Deno, 1985; Good & Kaminski, 2002; Wiederholt & Bryant, 2001). These tests are available for reliably measuring *components* of oral reading fluency rather than fluency as a whole construct. Recent reviews, however, have recommended measuring all components of fluency as part of a reading assessment program (Dowhower; 1991; Fuchs et al., 2001; Hudson et al., 2005; Hudson et al., 2009; Kuhn et al., 2010). However, they also describe the difficulty of finding reliable and valid assessments that incorporate all three fluency components.

Some scales have been designed to measure fluency as a whole construct (Allington, 1983; Pinnell et al., 1995; Zutell & Rasinski, 1991). However, brief analyses of these scales (above) reveal several flaws in the scales as they currently exist. First, there is a paucity of research regarding the reliability of scale scores when raters only have limited training. Second, the scales were all developed based on theoretical frameworks, but creators did not conduct empirical examinations of the individual fluency components to determine the degree to which accuracy, reading rate, and the various features of oral reading prosody contributed to fluency as a whole. Third, oral reading prosody is sometimes simply referred to as expression, and subjective descriptions of a reader's *expressiveness* often accompany fluency rating scales without specifically identifying the prosodic features to be focused on. For prosody to be an effective indicator of fluency, the components which have been shown to actually relate to reading ability and specifically, reading fluency, should be examined before inclusion in a rating scale (prosodic features have been shown to be differentially related to reading fluency and comprehension in empirical studies, e.g., Miller & Schwanenflugel, 2006; 2008).

These gaps in existing fluency assessments call for two broad steps to be taken: a theoretically and empirically based scale should be developed for assessing children's oral reading fluency as a whole construct, and the new scale should be rigorously tested and revised using methods that can provide evidence for a sound validity argument. My validation of the new scale will be based around the following goals for the scale's development:

1. Expressiveness components of an oral reading fluency scale will be grounded in the prosodic structure of children's oral reading, which can be measured spectrographically.

2. The scale will be consistent with current definitions, research, and theory in children's reading fluency and should be useful for its assessment.

3. Experts in children's reading will be able to use the scale with good inter-rater reliability.

4. Scale ratings of children's oral reading will correspond with spectrographic measures of children's prosody.

5. Scale ratings of children’s oral reading will correlate strongly with measures of children’s reading rate and accuracy and moderately with measures of reading comprehension.

Gathering the data, developing the scale, and performing the analyses to respond to these goals should provide the necessary evidence to make an argument regarding the validity of the new fluency scale according to Kane’s (1992) argument-based approach to validation.

Chapter 2: Study 1

The goal of Study 1 was to develop and pilot a spectrographically-grounded, valid rating scale of oral reading expression. This study proceeded in two phases. The goal of the first phase was to determine structure of the prosodic features of oral reading expression that needed to be captured to distinguish fluent from less fluent readers. Findings from prior studies on children's oral reading prosody (Benjamin & Schwanenflugel, 2010; Miller & Schwanenflugel, 2008) served as the basis from which the scale was developed. In particular, I focused on the prosodic feature data extracted by Benjamin and Schwanenflugel (2010) from children's oral readings that distinguished fluent from less fluent readers. In that study, recordings of children's oral readings of passages slightly below and above grade level were analyzed. In this study, I further analyzed the data to derive the structure of reading prosody and identify potential distinguishing features that needed to be captured by the scale. From this information, I developed a pilot version of the Comprehensive Oral Reading Fluency Scale. The goal of the second phase was to determine the validity and reliability of the scale. In this phase, I selected a subset of these oral readings to serve as the target recordings that would be evaluated by three experts (including me) using this initial scale. In phase 2, I evaluated the scale for issues of validity and reliability and determined changes that might need to be made to improve the scale. The basic strategy used in this study is depicted in Figure 2.1.

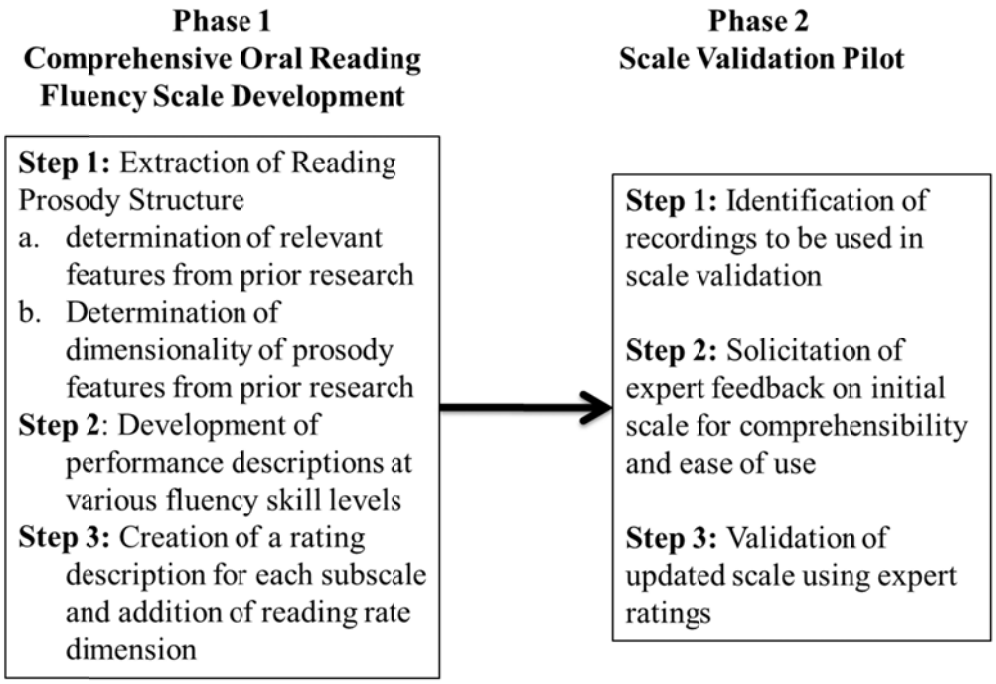
Phase 1
Comprehensive Oral Reading Fluency Scale Development
Step 1: Extraction of Reading Prosody Structure
a. determination of relevant features from prior research
b. Determination of dimensionality of prosody features from prior research
Step 2: Development of performance descriptions at various fluency skill levels
Step 3: Creation of a rating description for each subscale and addition of reading rate dimension
Phase 2
Scale Validation Pilot
Step 1: Identification of recordings to be used in scale validation
Step 2: Solicitation of expert feedback on initial scale for comprehensibility and ease of use
Step 3: Validation of updated scale using expert ratings

Figure 2.1. General strategy used in the development and pilot testing of the scale.

Method

Participants. Phase 1 used data from all 90 children's recordings from Benjamin and Schwanenflugel (2010). As reported, these children were all second-grade students (57% female, 43% male; mean age = 8 years 2 months, *SD* = 4 months; 50% European American, 32% African American, 10% Hispanic American, 4% Asian American, 2% other, and 1% unknown ethnicity) attending schools in either Georgia (83%) or New Jersey (17%). None were receiving special services for dual-language learners and all were able to read the "difficult" targeted passage from the Gray Oral Reading Test (1992, 2001).

Phase 2 used 59 recordings from this larger data set, such that the children were similar in demographics to the larger sample (53% female; mean age = 8 years, 2 months, *SD* = 4 months; 31% African American, 51% European American, 10% Hispanic

American, 5% Asian American, 2% other, and 1% of unknown ethnicity), attending schools Georgia (81%) or New Jersey (19%). These recordings were selected because they were relatively free of background noise. Chi-square goodness-of-fit tests revealed no significant demographic differences between the sub-sample and the larger set of participants from phase 1: sex χ^2 (1, *n* = 59) = .48, *p* = .489; ethnicity χ^2 (5, *n* = 59) = .53, *p* = .991; site χ^2 (1, *n* = 59) = .11, *p* = .737.

Three adults with expertise in assessment of children's reading fluency were selected to participate as *expert* raters. Expert raters consisted of individuals with doctoral level training specializing in the development of reading skill. Expert raters (excluding the author) were chosen for their availability and willingness to participate as well as a record of research publications in the field of oral reading fluency and in the larger field of the development of reading skill in general. One expert was a reading educator with over 15 years of experience training educators in the instruction and assessment the development of reading. The other was an educational psychologist with over 25 years of experience training educators in language and cognitive developmental factors related to schooling. The third (the current author) was a doctoral student and former English teacher with extensive knowledge of reading fluency and several research publications on the topic.

General assessments. The Benjamin and Schwanenflugel (2010) data set contained children's reading skill assessments administered during the spring term of second grade as well as reading prosody data. Children were administered the Gray Oral Reading Test (1992, 2001) to assess students' skill in reading connected text, and the Test of Word Reading Efficiency (TOWRE; 1999) Sight Word Efficiency to measure the automaticity with which children would be able to read sight words presented in a list. They were also administered the Reading Comprehension subtest of the Wechsler Individual Achievement Test (WIAT; 1992) to provide an indicator of children's general reading comprehension skills. All tests had previously reported reliability estimates ranging between .90 - .97 and validity estimates ranging between .39 and .94. All assessments were administered following instructions described in the test manual and all testers had been trained to a standard of 100% agreement with a trained school psychologist.

Extraction of reading prosody structure. For phase 1, prosodic measurements were taken from the spectrographic measures of prosody extracted from children's oral readings of an easy and difficult passage of the Gray Oral Reading Test by Benjamin and Schwanenflugel (2010). These prosodic measurements were carried out on two selected target passages from the GORT. The "easy" passage was passage 1 from the GORT-3 or passage 3 (the same passage) from GORT-4 and the "difficult" was passage 3 of the GORT-3 or passage 6 (the same passage) from the GORT-4, henceforth the difficult passage. Our own readability analyses averaging the Flesch-Kincaid Grade-Level Formula (Flesch, 1948) and the Spache Readability index (Spache, 1953) yielded an average estimated grade level of 1.97 for the easy passage, somewhat below the grade level of children in the study, and 3.79 grade level for the difficult passage, a good bit above the grade level of children in the study. A Lexile analysis found that the passages were appropriate for readers with early 2nd grade reading skills and mid to upper 3rd grade reading skills for the easy and difficult passages, respectively. For the spectrographic analysis, background interference was reduced from the oral reading files using noise reduction procedures, and prosodic analysis was carried out using Praat v.5.0.38. Praat is a free software program that is used to analyze, synthesize, and manipulate digital speech data (Boersma & Weenink, 2008).

Prosodic features used in determining the structure and organization of the reading prosody scale were those used by Benjamin and Schwanenflugel (2010) and had been found to relate significantly to other aspects of reading fluency. The following prosodic variables were used in the initial development of the scale: number of pausal intrusions (recorded as a ratio of pauses to possible pauses between words), number of pauses which cannot be grammatically justified, sentence-final pitch declination for declaratives, and pitch contour.

Pausal intrusions were measured by taking the participant's total number of within-sentence pauses divided by the total number of spaces between words. These pauses were counted by isolating and measuring the temporal space between words within a sentence. If the pause was 100 milliseconds or greater, it counted as one pausal intrusion (Benjamin & Schwanenflugel, 2010; Miller & Schwanenflugel, 2006; 2008). Prior studies have found that a measure of pause frequency within sentences is strongly related to other reading skills like automaticity and comprehension (Benjamin & Schwanenflugel, 2010; Clay & Imlach, 1971; Dowhower, 1987) Henceforth, pausal intrusion measurement from the text is referred to as passage "pause ratio."

Both Dowhower (1987) and Miller & Schwanenflugel (2006) showed that appropriateness of phrasing can impact a child's overall reading skill. That is, better readers tend to pause more appropriately than poor readers. Thus, Benjamin & Schwanenflugel (2010) introduced a measure of ungrammatical pausing, that is, the number of pauses within a sentence which cannot be grammatically justified. This feature was measured as the ratio of a participant's total within-sentence pauses which could not be explained by major shifts or cues within a sentence. Such shifts included the introduction of a new clause and cues included commas separating items in a list within a sentence. This ungrammatical pausing measure is henceforth referred to as passage "ungrammatical pause ratio."

Several earlier studies have found that betters readers drop their pitch more sharply at the ends of declarative sentences than struggling readers (Benjamin & Schwanenflugel, 2010; Clay & Imlach, 1971; Dowhower, 1987; Miller & Schwanenflugel 2008; Schwanenflugel et al., 2004). Sentence-final pitch (fundamental frequency measured in hertz) declination for declaratives was measured by isolating the target area on the spectrograph and measuring the pitch change in hertz from the final pitch peak to the end of the sentence. Magnitude of pitch declination was determined by subtracting the final pitch peak from the peak fundamental frequency. The mean difference in pitch across sentences was used as the indicator of sentence-final pitch declination. In some cases, "creaky voice" was observed in recording. Creaky voice is a result of irregular vocal fold vibration and can occur at any pitch. Thus, end-of-sentence prosody indicating creaky voice was not included as data (i.e., scored as missing), as this is generally not considered a valid indicator of pitch. In these cases, sentence-final pitch declination was based on the remaining sentences in selection. Henceforth, averaged sentence-final pitch change measurement from the text is referred to as passage "sentence-final pitch change."

Because the measure of intonation contour used by Benjamin & Schwanenflugel (2010; see also Miller & Schwanenflugel, 2006, 2008; Schwanenflugel et al., 2004) requires comparing children's intonation to adult intonation and thereby requires collecting adult oral reading samples to compute, I developed another measure of intonation contour from the Benjamin and Schwanenflugel (2010) recordings for exploratory purposes based on the variation that is expected in children's intonation while reading aloud (Snow & Coots, 1981) and the likelihood of such a measure providing a reliable indicator of pitch contours (Bolaños, Cole, Ward, Tindal, & Schwanenflugel, 2012; Cowie et al., 2002). The average pitch for the vocalic nucleus of each word was measured and a standard deviation obtained for each sentence. The standard deviations were averaged across sentences, resulting in a mean of the standard deviations of intonation for each child. This index of each child's intonation contour from the text is henceforth referred to as "pitch SD."

Development of the Comprehensive Oral Reading Fluency Scale.

Determination of scale dimensionality. As indicated in Figure 2.1, the second step for my Phase 1 goal was to determine how many dimensions the reading prosody component of the scale should have to ground the scale in spectrographically-derived measures for reading prosody. A principal components exploratory factor analysis was carried out using spectrographic prosody measurements from the full Benjamin and Schwanenflugel (2010) data set (*N* = 90) using prosody measurements from both the easy and the difficult stories. As stated above, an alternative measure of intonation contour was chosen for the present study to replace the adult-like intonation contour measure which required gathering adult prosody data.

Standardized scores of the prosody measurements were used to control for scaling differences among the measures (i.e., Hz versus ms pause). Varimax with Kaiser normalization rotation method was used. A principal components exploratory factor analysis allowed for each of the measured prosody variables to load onto as many components as determined to exist among the variables. As can be seen in Table 2.1, all communalities exceeded the common rule of thumb of .40.

Table 2.1

Principal Component Analysis Communalities

Communalities		
Reading Prosody Feature	Initial	Extraction
Easy Story Sentence-final Pitch	1.000	.680
Difficult Story Sentence-final Pitch	1.000	.509
Easy Story Pause Ratio	1.000	.704
Difficult Story Pause Ratio	1.000	.799
Easy Story Pitch SD	1.000	.863
Difficult Story Pitch SD	1.000	.788
Easy Story Ungrammatical Pause Ratio	1.000	.465
Difficult Story Ungrammatical Pause Ratio	1.000	.625

Note: Principal Component Analysis Extraction Method.

Only two factors had eigenvalues greater than 1.0 and examination of the scree plot confirmed that a two-factor solution would be the most appropriate. The two-factor solution accounted for 67.92% of the variance in the data and yielded two underlying components in reading prosody, with the first accounting for 37.90% of the variance and the second 30.02%. These factors were deemed interpretable as potentially reflecting two distinguishable aspects of reading prosody, *expressive intonation* and *natural pausing*. The loadings for the rotated solution can be found in Table 2.2. As can be seen in the table below, all features associated pitch changes loaded positively on the expressive intonation dimension, regardless of story difficulty. However, as indicated by the correlation between the factors, pausing was not completely independent of pitch changes. Pauses between words loaded negatively (albeit weakly) on the expressive intonation dimension, such that when children made larger pitch changes they also tended to pause less between words. For the natural pausing dimension, all prosodic features associated with pauses between words loaded positively on this dimension regardless of story difficulty. The factor analysis solution was shared with an expert on reading prosody who confirmed the reasonableness of this interpretation of the solution components. Correlations between all prosody variables can be found in Appendix A.

Table 2.2

Principal Component Analysis Rotated Component Matrix

Rotated Component Matrix[a]		
Reading Prosody Feature	Component	
	Expressive Intonation	Natural Pausing
Easy Story Pitch SD	.901	-.227
Difficult Story Pitch SD	.837	-.295
Easy Story Sentence-final Pitch	.815	-.125
Difficult Story Sentence-final Pitch	.708	-.092
Easy Story Pause Ratio	-.281	.791
Difficult Story Pause Ratio	-.417	.790
Easy Story Ungrammatical Pause	.097	.675
Difficult Story Ungrammatical Pause	-.304	.730

Note: Extraction Method: Principal Component Analysis with Varimax Rotation and Kaiser Normalization.

Development of performance descriptions for oral reading expression dimensions. Because the variable related to overall pause ratio (i.e., the ratio of actual pauses to the number of potential pauses) also loaded marginally onto the component dominated by expressive intonation features of prosody, performance descriptions of expressive reading also incorporated descriptions of expressive pausing, or pausing that enhances meaningful expression. Performance descriptions of *expressive intonation* directly reflected the findings from recent studies of children's expressive reading (Benjamin & Schwanenflugel, 2010; Miller & Schwanenflugel 2006; 2008; Schwanenflugel et al., 2004). That is, skilled readers tend to drop their pitch more at the ends of declarative sentences than less skilled readers do. Skilled readers vary their pitch more throughout a sentence than less skilled readers do.

Earlier scales of oral reading fluency (e.g., Pinnell et al., 1995; Rasinski et al., 2009) seemed to indicate that four performance levels within each dimension was a number with which reliable ratings could be achieved. While Allington's scale (Allington, 1983) seems to indicate that six levels may be appropriate, there is little evidence of reliable use of the scale in the literature (cf., Young et al., 1996). After conducting the principal components analysis and listening to the sound files for the dimensions described above, I decided that four performance levels of each dimension would be most appropriate for the current scale and would be distinguishable by raters (See Appendix B for the first version of the Comprehensive Oral Reading Fluency Scale and Appendix C for the final draft of the scale used for Study 1).

To develop the rating system using four performance levels, the Benjamin and Schwanenflugel (2010) prosody measurement data was examined using scatterplots to see trends in prosodic behavior as automaticity and comprehension changed. Additionally, line graphs had been reported in several studies (Benjamin & Schwanenflugel, 2010; Miller & Schwanenflugel, 2006; Schwanenflugel et al., 2004) in which children had been divided into skill groups (based on their reading automaticity) and the prosodic variables plotted by skill group. Finally, the actual prosodic measurements of students were carefully examined. For example, close examination of the data showed that not only did skilled readers tend to vary their pitch more than struggling readers, but they varied their pitch in more appropriately than struggling readers. The performance descriptions of the initial versions of the Comprehension Oral Reading Fluency Scale, then, were designed to reflect children's actual prosodic capabilities when reading aloud.

Using a 4-level rating system for the *expressive intonation* dimension, level 4 described reading characterized by prosody that is associated with high levels of comprehension (Benjamin & Schwanenflugel, 2010; Miller & Schwanenflugel, 2006; 2008). Students reading at this level would consistently and appropriately vary their intonation to match the meaning of a sentence; their pausing would match the syntax of the text; and their sentence-final pitch would be appropriate. In contrast, level 1 described reading characterized by prosody associated with poor comprehension. Students reading at this level would consistently read with flat or inappropriate intonation and would fail to vary their pitch to match sentence-final punctuation. Levels 2 and 3 demonstrated variations in degree. A student reading at level 3, for example, would exhibit most of the characteristics of level 4 reading, but might occasionally read with inappropriate intonation or fail to mark the end of a sentence with appropriate pitch. A student reading at level 2 might make some attempts at reading with appropriate prosody, or might occasionally mark the end of a sentence with an appropriate pitch change, but reading would frequently fail to exhibit these characteristics.

Performance descriptions for the *natural pausing* dimension reflected findings from earlier studies which found that children with greater reading skill pause less between sentences and have fewer pauses within sentences than less skilled children (e.g., Benjamin & Schwanenflugel, 2010; Miller & Schwanenflugel, 2006). Benjamin and Schwanenflugel (2010) also found that while more skilled readers may pause some within sentences, those pauses tend to reflect natural grammatical breaks in the text (e.g., a skilled reader may pause between clauses within a sentence, or while reading a list). Thus, performance descriptions of this dimension included descriptions of both within- and between-sentence pausing as well as evaluation of the appropriate placement of pauses while reading aloud.

Using a 4-level rating system for the *natural pausing* dimension, level 4 described reading characterized by pausing patterns that are associated with high levels of comprehension (Benjamin & Schwanenflugel, 2010; Miller & Schwanenflugel, 2006; 2008). At this level of reading any pauses that exist would only appear at appropriate places in the text—e.g., between sentences or at natural syntactic or semantic shifts within a sentence—and these pauses would be brief. In contrast, level 1 described reading characterized by consistent effortful and broken reading. Pauses would be numerous and would not necessarily be consistent with the semantic and syntactic structure of the text. Levels 2 and 3 demonstrated variations in degree. A student reading at level 3, for example, would exhibit most of the characteristics of level 4 reading, but might momentarily have a lengthier pause or pause at an inappropriate place in the text. A student reading at level 2 might have bursts of connected reading but would pause frequently and non-systematically. Pausing between sentences would also be lengthy and would interrupt the flow of the text.

Performance descriptions for rate and accuracy dimension of reading fluency. Reading fluency has traditionally been measured by simply calculating the number of words a child can read correctly in one minute (Hasbrouck & Tindal, 2006). Because this method serves to quickly and accurately measure a child's ability to read connected text with automaticity, it is unlikely that replacing such an objective measure with more qualitative performance descriptions of rate and accuracy would be an improvement. Thus, the *rate and accuracy* dimension of the scale was designed to reflect Hasbrouck and Tindal's (2006) WCPM norms for children at the appropriate grade level and season of the year. This method allows the scale to be customizable for children at various grade levels since the WCPM norms for a second grade child during the winter differ from the norms for a third grade child in the fall, for example. Performance descriptions, then, are based on the Hasbrouck and Tindal published quartiles.

Benjamin and Schwanenflugel (2010) examined the relative strength of prosody's predictive relationship with comprehension vs. rate and accuracy's predictive relationship with comprehension. They found that when rate and accuracy were entered first in a two-step regression, prosody accounted for an additional 5.5% of variance in comprehension scores. When prosody was entered first, rate and accuracy accounted for an additional 6.2% of the variance in comprehension scores. Prosody and rate and accuracy were quite similar in their predictive powers. Thus, it was determined that prosody should be weighted equally with rate and accuracy in the Comprehensive Oral Reading Fluency Scale. The *Rate and Accuracy* dimension is weighted, then, at 50% of the comprehensive oral reading fluency score. Because there are two rated sub-dimensions within the oral reading fluency expression dimension, the scale scores for the rate and accuracy dimension should range from 2-8 while scale scores for each of the expression dimensions range from 1-4, allowing students to earn up to 8 total points for expression and up to 8 total points for rate and accuracy. Thus, a child with WCPM score in the lowest quartile earns a score of 2 while a child in the highest quartile earns a score of 8.

Expert feedback. Expert raters were sent drafts of the scale while it was being developed. They were given the opportunity to openly respond to the scale and provide suggestions for changes based on their expertise. Based on expert feedback, changes were made to the wording of the performance descriptions and the clarity of instructions. However, the original dimensionality of the scale was retained as well as the number of performance levels. Once the scale was revised, expert raters began rating children's oral readings.

Expert rating procedures. Experts were sent a CD with 59 children's readings of the easy passage only from Benjamin and Schwanenflugel (2010) and a copy of the new rating scale. The 59 readings used in the study were selected based on the clarity of the recordings. I determined that raters would need to have recordings relatively free of background noise in order to conduct ratings without too much frustration. Thus, these 59 readings were selected based on my subjective judgment of the recordings' quality and clarity. While Benjamin and Schwanenflugel used two passages, an easy and a difficult passage, comparable oral reading fluency rating scale developers have used their scales with passages roughly reflecting the grade level of their readers (Daane et al., 2005; Pinnell et al, 1995; Rasinski, 2004). Additionally, phase 2 of the current study was designed to simply serve as a pilot study for the initial development and testing of the Comprehensive Oral Reading Fluency Scale. Issues of text level appropriateness were addressed in study 2. Instructions, developed by Benjamin, were sent to raters as well (see the final scale for Study 1 in Appendix C). Raters initially rated 11 oral readings to determine if sufficient reliability could be obtained without formal training. No substantive changes were made to the scale following this initial rating sequence, so initial ratings were included in the total ratings for this sample of participants. The expert raters independently conducted ratings in a quiet and isolated area free from distraction and background noise. All raters rated the recordings without knowledge of children's performance on standardized assessments of reading skills. Experts rated a total of 59 children's oral reading recordings of the passage.

Results

Descriptive statistics of prosody measurements and standardized assessments. Prior to performing statistical analyses, data were analyzed for outliers using standard scores for all variables. One child earned exceptionally high scores on the GORT and the TOWRE, and also had an exceptionally high pitch SD score. This child fit the sample by age and grade level, however, and was retained. All standardized test scores and prosody measurements were examined for mean, range, SD, skew, and kurtosis. All values were deemed initially acceptable based on the types of analyses being performed (see Table 2.3).

Table 2.3

Descriptive Statistics of Standardized Tests and Prosody Measures

	Minimum	Maximum	*M*	*Mdn*	*SD*
GORT-rate	5	20	11.59	12.00	3.05
WIAT-RC	81	146	111.14	110.00	14.69
TOWRE	89	145	113.31	112.50	12.03
Pause ratio	0	.77	.20	.14	.18
Intersentential pause length	0	990	420	377	221.76
Ungrammatical pause ratio	0	1	.67	.78	.36
Sentence-final pitch change	-24.22	106.15	39.44	39.32	29.44
Pitch SD	6.89	52.04	21.64	20.56	10.22

Note: *n* = 59; GORT-rate = Gray Oral Reading Test, standardized rate & accuracy measurement; WIAT-RC = Wechsler Individual Achievement Test, reading comprehension subtest; TOWRE = Test of Word Reading Efficiency, site-word efficiency subtest. The normed mean for the GORT-rate is 10 (*SD* = 3). The normed mean for both the WIAT-RC and the TOWRE is 100 (*SD* = 15).

Interrater reliability. Interrater reliability was examined using two methods: 1) rater agreement percentages and 2) intraclass correlations. Intraclass correlation coefficients (ICCs; as opposed to *inter*class correlations, e.g., Pearson r) are generally obtained when comparisons of scores within participants on the same assessment are desirable (McGraw & Wong, 1996). Conventions for qualitatively describing the strength of ICCs are drawn from descriptions used for Kappas (Landis & Koch, 1977), such that an ICC of less than .40 is considered “poor,” between .40 and .59 is considered “moderate,” between .60 and .79 is considered “substantial,” and an ICC above .80 is “outstanding.” Henceforth, terminology used to describe ICCs will reflect these descriptors. In the present study, two raters measured WCPM for each of 60 participants and also rated each participant using the Comprehensive Oral Reading Fluency Scale. Based on the criteria set forth by Shrout and Fleiss (1979), for the present study in which two raters each rated the entire sample of participants, an ICC obtained through a Participant X Raters two-way random effects ANOVA is the appropriate method of analysis to use and is the method of ICC used throughout study 1. All intraclass correlation analyses conducted in this study use absolute agreement rather than simply consistency as the standard for comparisons and all ICCs reported are based on a single-measure rather than average-measure analysis. Descriptive statistics of ratings, including the mean, median, mode, *SD*, and range of each rater are found in Table 2.4.

Table 2.4

Descriptive Statistics of Ratings Using the Comprehensive Oral Reading Fluency Scale

	Minimum	Maximum	*M*	*Mdn*	Mode	*SD*
Rater 1 Intonation	1	4	3	3	4	0.95
Rater 2 Intonation	1	4	2.56	2	2	0.73
Rater 3 Intonation	1	4	3.02	3	3	0.82
Rater 1 Pausing	1	4	3	3	3	0.87
Rater 2 Pausing	1	4	2.81	3	3	0.80
Rater 3 Pausing	1	4	2.92	3	3	0.79
Rater 1 Total Expression	2	8	6	6	7	1.58
Rater 2 Total Expression	2	8	5.37	6	6	1.31
Rater 3 Total Expression	3	8	5.97	6	6	1.47
Rater 1 Total Score	5	16	13.12	14	15	2.94
Rater 2 Total Score	5	16	12.46	14	14	2.81
Rater 3 Total Score	5	16	13.08	14	14	2.87

Note: *n* = 59.

WCPM and the Rate and accuracy dimension. Descriptive statistics for students' WCPM scores can be found in Table 2.5. WCPM data resembled a Normal distribution. Percent agreement was used to examine agreement among the ratings of the Rate and Accuracy dimension. Agreement among raters on the rate and accuracy dimension was high, as would be expected on a rating dimension that is based on the objective measurement of WCPM. Percent agreement across all three raters for all ratings can be found in Table 2.6. An intraclass correlation coefficient was obtained to determine reliability among students' WCPM among all three raters. The ICC for rate and accuracy was outstanding, ICC = .99, *F*(58, 116) = 363.38, *p* < .001.

Table 2.5

Descriptive Statistics of WCPM as Assigned by Raters

	Minimum	Maximum	*M*	*SD*
Rater 1 WCPM	50	240	145.81	43.78
Rater 2 WCPM	48	248	142.15	44.10
Rater 3 WCPM	49	240	144.85	43.25

Note: *n* = 59.

Table 2.6

Percent Agreement Across All Raters

	Exact Agreement			Adjacent Agreement		
	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
			Rate & Accuracy			
Rater 1	--	98	100	--	100	100
Rater 2	--	--	98	--	--	98
			Expressive Intonation			
Rater 1	--	39	69	--	90	98
Rater 2	--	--	37	--	--	97
			Natural Pausing			
Rater 1	--	54	66	--	100	98
Rater 2	--	--	66	--	--	100
			Total Expression			
Rater 1	--	39	42	--	71	93
Rater 2	--	--	27	--	--	86
			Total Scale Score			
Rater 1	--	37	42	--	69	93
Rater 2	--	--	27	--	--	86

Note: *n* = 59

Expression dimensions. Percent agreement analyses were conducted to determine the level of interrater agreement among the three raters. Ratings for *Expressive Intonation* and *Natural Pausing* each ranged from one to four. Due to the limited range of scores on the individual dimensions as well as the need for agreement on the ratings rather than simply consistency, the most significant analysis of these ratings is that of percent agreement (see Table 2.6). Agreement among raters on the individual expressive intonation and natural pausing dimensions was low to moderate with greater agreement among ratings of pausing behavior vs. ratings of intonation. Adjacent agreement was high, however, but on a scale with scores ranging from one to four, such high adjacent agreement would be expected. Adjacent agreement for *Expressive Intonation* and *Natural Pausing* ratings is comparable to adjacent agreement among highly trained raters using the NAEP scale (Daane et al., 2005; Pinnell et al, 1995). Intraclass correlations were conducted for expressive intonation and natural pausing ratings as an additional measure of interrater reliability. The ICC for expressive intonation was moderate, ICC = .56, *F*(58, 116) = 5.74, *p* < .001. The ICC for natural pausing was substantial, ICC = .71, *F*(58, 116) = 8.54, *p* < .001. Improvements to reliability for the expressive intonation dimension will likely result from further modifications to the scale.

Total expression scores. Percent agreement was analyzed for total expression scores with the goal commonly being that adjacent agreement of 90% or higher among highly trained raters might be achieved. For raters without formal training, however, the 90% adjacent agreement may not be realistic. Nonetheless, as Table 2.6 illustrates, this goal was achieved between raters 1 and 3. An intraclass correlation between all three raters was conducted for the total expression scores. The ICC among all three raters' total expression scores was analyzed, as the total expression scores ranged from two to eight and could be considered a continuous variable for statistical purposes. Among the three raters, the ICC using absolute agreement as the standard was substantial, ICC = .75, *F*(58,

116) = 12.22, *p* < .001. After examining the ratings and getting feedback from raters, it became clear that rater 2 had difficulty objectively rating student readings according to the scale's performance descriptions, and this rater communicated that she did not believe that her ratings accurately reflected students' performance on the scale per se. Rather, her experience with other rating scales and reading assessments significantly affected her judgment and her ability to give students high scores even though they may have performed well. Essentially, the scale descriptions at the high end, in particular, did not reflect what she believed a fluent reader should sound like. Further, she felt her extensive experience instructing reading specialists to use similar scales had interfered with her ability to use this scale. This self-evaluation was consistent with analysis of ratings patterns as seen in Table 2.4. Rater 2 had a lower mean than other raters across dimensions. Thus, a second ICC was obtained using total expression ratings from only raters 1 and 3. This coefficient was outstanding, ICC = .83, *F*(58, 58) = 10.75, *p* < .001.

Comprehensive oral reading fluency scores. Percent agreement was analyzed for total scale ratings with the goal being at least 90 percent agreement of +/- 2 points (cf. Rasinski et al., 2009). All raters achieved 100% agreement. Percentages of exact and adjacent agreement among raters' comprehensive oral reading fluency scores can be found in Table 2.6. An intraclass correlation between all three raters was conducted for the comprehensive oral reading fluency scores. The ICC among all three raters' scores was analyzed, as the comprehensive oral reading fluency scores ranged from four to sixteen and could be considered continuous variables for statistical purposes. Among the three raters, the ICC = .93, *F*(58, 116) = 53.66, *p* < .001. Thus, even though rater 2 had expressed doubts about her use of the scale, the ICC between raters was quite good.

Relationship between expression ratings and prosody measurements. Correlational relationships between ratings and prosody variables were examined to investigate the scale's potential as an assessment that can be used to roughly gauge children's prosody. To simplify interpretation within the study and comparisons across studies, the three raters' ratings were averaged for each dimension; i.e., if rater 1 gave a child a rating of 3 for appropriate intonation, rater 2 gave the child a 4, and rater 3 gave the child a 3, then a score of 3.3 is used to correlate the appropriate intonation dimension with prosody variables. All correlations of individual raters' expression ratings with the prosody variables can be found in Appendix D.

Expressive intonation. Ratings were averaged to examine the correlational relationships between the ratings and related prosody measurements: sentence-final pitch change, pitch SD, and pause ratio. These three variables were found to load onto the same factor in the principal components analysis carried out in phase 1. Using a linear Pearson correlation, mean expressive intonation ratings correlated strongly with all variables: sentence-final pitch (r = .56, p < .001), pause ratio (r = -.67, p < .001), and pitch SD (r = .66, p < .001).

Natural pausing. Ratings were averaged to examine the correlational relationships between the ratings and related prosody measurements: pause ratio and ungrammatical pause. A third pause variable was measured and analyzed as a pilot variable: intersentential pause length. As noted earlier, pause ratio and ungrammatical pause were variables identified by Benjamin and Schwanenflugel (2010) that loaded onto the same factor in the principal components analysis carried out in relation to phase 1. The last variable, intersentential pause length, was found to be a useful predictor of oral reading skill in many prior studies (e.g., Miller & Schwanenflugel, 2006; Schwanenflugel et al., 2004). Based on such prior research, performance descriptions of the *Natural Pausing* scale dimension mention intersentential, or between-sentence, pausing. Intersentential pause length was measured by selecting the space between the final word of a sentence and the beginning of the first word of the following sentence; that space was selected on the spectrograph and measured in milliseconds, ranging from a minimum of 100 ms to a maximum of 3000 ms, as dictated by the general testing protocol required by the GORT. The average length of each participant's inter-sentential pauses was recorded. Henceforth, intersentential pause length from the text will be referred to as passage "intersentential pause length." Descriptive statistics for children's Intersentential pause length can be found in Table 2.1.

Using a linear Pearson correlation, mean natural pausing ratings correlated strongly with pause ratio ($r = -.82$, $p < .001$) and intersentential pause length ($r = -.63$, $p < .001$). The correlation between mean ratings and ungrammatical pause was moderate but much lower than other correlations ($r = -.33$, $p = .010$). Note that correlations between ungrammatical pause and other pause variables had been reported as low in prior research (see Benjamin & Schwanenflugel, 2010). Ungrammatical pause may be a less valid indicator of prosodic reading than other pause variables. The ungrammatical pause length correlation was significantly lower than those of both pause ratio ($t = 5.16$, $p < .001$) and intersentential pause length ($t = 2.22$, $p = .02$). Regardless, ratings of natural pausing did seem to correspond reasonably well in general with actual pause measurements obtained from spectrographic analyses.

Total expression. As with appropriate intonation and natural pausing ratings, total expression ratings were averaged across raters to examine their correspondence to all prosody measurements: sentence-final pitch ($r = .54$, $p < .001$), pitch SD ($r = .62$, $p < .001$), pause ratio ($r = -.81$, $p < .001$), ungrammatical pause ($r = -.26$, $p = .044$), and intersentential pause length ($r = -.55$, $p < .001$). All correlations were moderate to high except for the correlation between the mean total expression rating and ungrammatical pause, which was much lower than correlations among the other variables. This result in conjunction with the low correlations between ungrammatical pausing and the other prosodic pause variables (see Benjamin & Schwanenflugel, 2010) suggests that at texts of this level of difficulty, ungrammatical pausing may not play a practically significant role in children’s reading skill. Benjamin & Schwanenflugel (2010) found that good readers had significantly fewer ungrammatical pauses than poor readers when reading both easy and difficult texts. However, easy passage ungrammatical pausing correlated weakly with all other prosody variables, and ungrammatical pausing did not play a significant role in regression equations developed using prosody variables to predict automaticity and comprehension. Thus, while ungrammatical pausing was successful in distinguishing good from poor readers, it was not as robust as other prosodic variables.

Relationship between total rating scores and standardized assessments. Correlational relationships between ratings and traditional reading assessments were examined to investigate the scale’s potential as an assessment that can be used to roughly gauge children’s prosody. To simplify interpretation within the study and comparisons across studies ratings were averaged across raters. Correlations of individual raters’ total expression scores and comprehensive oral reading fluency scores with the standardized test scores, as well as correlations among standardized test scores, can be found in Appendix D.

Total expression scores. Correlational relationships were examined between total expression scores and standardized tests of text reading fluency (GORT-rate), word reading fluency (TOWRE), and reading comprehension (WIAT-RC). All correlations between total expression and standardized test scores were high: GORT-rate ($r = .86$, $p < .001$), TOWRE ($r = .74$, $p < .001$), and WIAT-RC ($r = .73$, $p < .001$). Results suggest a strong relationship between children's prosody and other reading skills.

Comprehensive oral reading fluency scores. Correlational relationships were examined between comprehensive oral reading fluency scores and standardized tests of text reading fluency (GORT-rate), word reading fluency (TOWRE), and reading comprehension (WIAT-RC). All correlations between comprehensive oral reading fluency scores and standardized test scores were high: GORT-rate ($r = .84$, $p < .001$), TOWRE ($r = .75$, $p < .001$), and WIAT-RC ($r = .69$, $p < .001$). Results suggest a strong relationship between children's performance on the Comprehensive Oral Reading Fluency Scale and their performance on more traditional tests of reading skill, particularly tests of automaticity. This relationship is expected since a 50% weight in the comprehensive scale score is given to the rate and accuracy rating.

Rater feedback. Raters were instructed to keep track of the time it took them to conduct ratings using the scale (rater burden) and were also asked to provide feedback regarding the instructions, performance descriptions, format, and general usefulness of the scale. After conducting ratings for study 1, rater provided significant helpful feedback. Raters reported an average burden of four to five minutes per rating. They suggested that this length of time is reasonable for researchers and expert raters who have access to recordings of children's reading. Regarding the scale instructions, raters suggested providing a comment regarding the goal of the scale as a criterion-referenced assessment. Many of those using the scale will have had experience with other fluency rating scales, so some direction should be provided about adhering to the performance descriptions alone when ratings children's reading. Also, it was suggested that repetitions and hesitations be addressed so that raters would know how whether or not to count these as errors.

Raters generally approved of the scale's format but provided several suggestions for changes to performance descriptions including the following: 1) consolidating some of the expressive intonation level 4 description to shorten it, 2) making some mention of possible but rare unexpected pausing even among the best of readers for the natural pausing dimension. It was suggested that some raters may be reluctant to give readers a natural pausing rating of 4 if the reader stumbles at all while reading. Raters asked that possible changes be made to the level 4 natural pausing description to reflect the possibility that a reader might make an occasional misstep. Finally, a rater suggested that perhaps changing the heading "expressive intonation" to "appropriate intonation" might help raters avoid the mistaken expectation that the most prosodic readers will read as if they were performing a reader's theater production. In sum, raters made suggestions that would only result in minor changes to the scale as a whole. Raters expressed that the scale appeared to be valid and useful for measuring children's oral reading fluency.

Chapter 3: Studies 2a and b

The goal of Study 2 was to test and refine the spectrographically-grounded oral reading fluency scale that had been developed and initially tested in Study 1. Study 2 was broken up into two sub-studies. In Study 2a I modified the scale and some methods based on the results of Study 1 and then tested the modified scale on a sample of participants. The results of Study 2a informed Study 2b, in which I made final modifications to the scale and conducted a final test. Modifications are discussed in greater detail in the Methods sections for each of these studies.

General Modifications to Methods Based on Results of Study 1

Study 1 was conducted using a largely pre-existing data set from Benjamin and Schwanenflugel (2010). Because Study 2 involved collecting new data, several changes were made based on results of Study 1 and the need to conduct a relatively authentic and robust validation study of the Comprehensive Oral Reading Fluency Scale. First, the ungrammatical pause variable used in Benjamin and Schwanenflugel as well as in Study 1 did not correlate strongly with other prosody variables or with scale results. Additionally, regression equations published in Benjamin and Schwanenflugel revealed that ungrammatical pausing played an insignificant role in predicting both automaticity and comprehension. Thus, I decided to remove ungrammatical pausing as a variable in Study 2. In contrast, while intersentential (between sentence) pause length was not used in the initial developmental phase of the scale in Study 1, it was piloted in phase 2 of Study 1 and was found to correlate strongly with other prosody variables, with scale scores, and with standardized tests. Because of its strong performance in prior studies as well (e.g., Clay & Imlach, 1971; Cowie et al., 2002; Miller & Schwanenflugel, 2006; 2008; Schwanenflugel et al., 2004), I decided to include the variable permanently in Study 2.

While a passage from the Gray Oral Reading Test (GORT) was used in phase 2 of Study 1, it seemed more appropriate in Study 2 to use passages that might more closely resemble passages used in classrooms and studies in which children's reading characteristics are examined in detail. Additionally, I wanted to be able to have raters rate passages based on an average one minute of reading, and the passage from the GORT was much less than one minute of reading for most children. Thus, passages from the Qualitative Reading Inventory (QRI-5) were selected because of their relative complexity, length, and authenticity based on consultation with three reading experts who each have numerous publications in the field of reading development and fluency as well as over 15 years of experience in reading research and assessment or combined experience in classroom reading instruction and research. Two passages of different difficulty levels from the QRI-5 were selected since the reading abilities of the children were not yet known and some research has suggested that prosodic measurements from more challenging texts may provide more information about a child's overall reading ability than prosody from easier texts (Benjamin & Schwanenflugel, 2010; see also Young & Bowers, 1995). Since the text used in Study 1 was somewhat leveled slightly below the participants' grade level (an early 2nd grade text was used with children who were at the end of 2nd grade) I wanted to see if the scale could provide consistent results even across texts that varied somewhat in difficulty.

In sum, this study was designed to evaluate the validity of the updated scale using relatively authentic texts of varying difficulty and a different and more representative participant population (as described below). Study 2a allowed me to test the Comprehensive Oral Reading Fluency Scale’s ability to accurately measure children’s reading fluency while at the same time testing raters’ ability to use the scale reliably. That study indicated that some further changes might need to be made to hopefully increase inter-rater reliability. So, modifications were made and an updated final test of the scale was carried out in Study 2b.

Study 2a Method

Participants. 120 third grade children from public schools in Georgia and New Jersey participated in the study (52% female; mean age = 9 years 4 months, SD = 4.8 months; 21% African American, 64% European American, 9% Hispanic, 3% Asian, 3% Other; 77% from Georgia, and 23 % from New Jersey). 42% of children were receiving free or reduced price school lunches. Only those children participating in regular education classrooms and not currently receiving English language support services were included in this study. Stratified random sampling by region was used to divide children into two sampling groups of 60 participants each for studies 2a and 2b. This was to ensure that locations were equally represented across samples. Teachers received six children’s books donated to their classroom libraries as thanks for their participation.

Two adults with experience in listening to recordings of children reading aloud and relative expertise in reading fluency were selected to participate as *expert* raters. Expert raters had at least 10 years of research experience on children’s reading fluency and at least 10 research publications on the topic. One of these experts participated in the previous study and the other was a new user of the scale. The expert raters (excluding the author) were chosen based on availability and willingness to participate, as well as expertise with other rating schemes of oral reading fluency.

General assessments and procedures. For children, formal reading assessments were administered during the end of the spring term of third grade. Children were administered components of an informal reading inventory in a standardized fashion and standardized tests of word reading efficiency and reading comprehension. Participants received the informal reading inventory assessments first and then the standardized word reading efficiency and comprehension tests. All assessments were carried out by trained testers. All children were tested individually in a quiet location in their schools.

Oral reading rate and accuracy. The Qualitative Reading Inventory, 5th edition (QRI-5; Leslie & Caldwell, 2011) is an informal reading inventory (IRI), which provides users with numerous assessment options for children at pre-primer reading levels through high school. Because the QRI-5 was not designed as a standardized assessment, users can choose which portions of the assessment are relevant for their students or participants. Oral reading rate and accuracy can be easily measured with the QRI-5 by having a student read a grade level passage aloud, timing the reading, and counting the number of deviations from print while reading. The number of words read correctly per minute (WCPM) is then calculated as the child's score. To obtain each child's grade level WCPM, children read *Where do People Live?*, and to obtain each child's above grade level WCPM, children read *Early Railroads*. The test manual (Leslie & Caldwell, 2011) reports inter-scorer reliability for recording miscues to be α = .99 when comparing persons without extensive training (e.g., undergraduates) to person with extensive training (i.e., reading teachers or specialists with masters degrees). Extensive training, then, is not necessary for accurate scoring of miscues. *Where do People Live?* is designated for the third grade level, and *Early Railroads* is designated for the fourth grade level. Several readability formulas and leveling systems reported in the QRI-5 manual (Leslie & Caldwell, 2011) indicated that these passages vary in difficulty. The QRI-5 manual also indicated that *People* is indexed at 500 Lexiles and contains 279 words. *Railroads* is indexed at 810 Lexiles and contains 297 words. While specific miscue analysis was not of interest for the present studies, interrater agreement for WCPM was high for both passages, $r = .99$, $p < .001$.

Word reading efficiency. To obtain an independent estimate of word reading efficiency, children were administered the Test of Word Reading Efficiency (TOWRE) Sight Word Efficiency subtest (1999), which assesses the number of real words correctly read from a list within 45 seconds. Children were assessed using TOWRE Form A; the subtest raw score was converted to a standard score based on age, as directed by the examiner's manual. Test–retest reliability calculated for children ages 6-9 years is reported as 0.97 in the manual. Furthermore, concurrent validity estimates reported in the manual have a coefficient of 0.92 for third-grade students. A 15% random subsample of the present participants was selected for inter-scorer reliability. Reliability was high, $r = .99$, $p < .001$.

Informal reading comprehension. The QRI-5 (Leslie & Caldwell, 2011) described above allows educators and researchers to test the comprehension of read passages by answering eight questions immediately following the student's reading of the passage. Four of the questions were "explicit" questions defined as questions where the answers can be identified directly in the passage, and four of the questions were "implicit" questions defined as questions where the answers can be formulated from clues in the passage. An item scored either correct or incorrect using the correct answers provided by the QRI-5. For the purpose of this study, students were not permitted to engage in "look backs" while responding to comprehension questions, and assistance from the examiner was not given.

The test manual (Leslie & Caldwell, 2011) reports inter-scorer reliability for both explicit and implicit comprehension questions to be $\alpha = .98$, very high when comparing persons without extensive training (e.g., undergraduates) to person with extensive training (i.e., reading teachers or specialists with masters degrees). Extensive training, then, was deemed unnecessary for accurate scoring of comprehension questions. A 15% random subsample of the present participants was selected for inter-scorer reliability. Reliability was moderate for both the level 3 and level 4 passages, respectively, $r = .81$, $p < .001$; $r = .87$, $p < .001$.

General reading comprehension. The Reading Comprehension subtest of the Wechsler Individual Achievement Test, 3rd edition (WIAT; Wechsler, 2009), was administered to obtain an independent measure of the students' reading comprehension skill. This subtest consisted of a series of printed passages with increasing difficulty, followed by a question presented and responded to orally. The subtest contained both literal and inferential comprehension question types. The children were instructed to read a passage, listen to the question presented by the examiner, and then respond orally in their own words. Procedures for administration were followed as described in the test manual. This test measures reading comprehension as children's ability to answer questions about the text, a skill that many teachers consider a key indicator of reading comprehension (Richardson, Anders, Tidwell, & Lloyd, 1991). The raw score, determined by the number of questions answered correctly, was converted to a standard score based on age, which then served as an indicator of reading comprehension skill, henceforth referred to as WIAT-RC. The test manual reports the split-half reliability coefficient for this age range (8-10 years) as a mean of 0.91 and the validity estimates compared with other reading comprehension tests fall between 0.74 and 0.79. A 15% random subsample of the present participants was selected for inter-scorer reliability. Reliability was high, $r = .93$, $p < .001$.

Reading prosody assessment and procedures. Prosodic measurements were carried out on the two target passages from the QRI-5. Because it is standard procedure in curriculum based measures (CBMs) as well as various fluency rating scales (e.g., Rasinski, 2004) to measure a child's performance simply based on one minute of reading aloud, only a portion of each text was selected for prosodic measurement and rating. Some research has shown that simply selecting the first minute of reading may artificially inflate a child's overall reading rate (Valencia et al., 2010); thus, portions of text from the middle of the narratives were selected for analysis and rating. However, only full sentences were considered in order to obtain as much prosodic information as possible on units of text.

Readings from the children were obtained using digital voice recorders. Additionally, a shareware version of the Audacity (version 1.3.14; Audacity Developer Team, 2011) digital audio editor was used to create individual WAV files for prosody analysis. Background interference was reduced using noise reduction procedures, and prosodic measurements were carried out using Praat (version 5.2.22; Boersma & Weenink, 2011), a free software program that analyzes, synthesizes, and manipulates digital sound and speech data.

Measurement of prosodic features. Prosodic features to measure were based on the following criteria: 1) features have been found to be significantly related to other aspects of reading fluency in previous research; 2) features may have been found to be significantly related to reading comprehension in previous research; 3) the ability to measure the features exists in the texts used for analysis. Based on these criteria, the following prosodic features were measured: intersentential pause length, number of pausal intrusions (recorded as a ratio of pauses to possible pauses between words), sentence-final pitch declination for declaratives, and intonation contour. Following is a brief description of the procedures used to measure each prosodic feature. More detailed procedures can be found in Appendix E.

Intersentential pause length was measured by selecting the space between the final word of a sentence and the beginning of the first word of the following sentence; that space was selected on the spectrograph and measured in milliseconds (ms) up to 3000 ms since procedures specified that children be instructed to move on after pausing for three seconds. The average length of each participant’s intersentential pauses was recorded. Henceforth, intersentential pause length from the grade level text will be referred to as “level 3 intersentential pause length” and from the above grade-level text as the “level 4 intersentential pause length.”

The number of pausal intrusions was measured as the participant’s total number of within-sentence pauses divided by the total number of spaces between words. The ratio was recorded as the number of pausal intrusions so that comparisons could be made between the grade-level and above grade-level passages. These pauses were counted by isolating and measuring the temporal space between words within a sentence, including hesitations pre-articulation as pausing. In most cases, however, the hesitation or pre-articulation only added to the length of an already existing pausal intrusion. If the pause was 100 milliseconds or greater, it counted as one pausal intrusion (Benjamin & Schwanenflugel, 2010; Miller & Schwanenflugel, 2006; 2008). Henceforth, pausal intrusion measurement from the grade-level text is referred to as "level 3 pause ratio," and from the above grade-level text as "level 4 pause ratio."

Sentence-final pitch declination for declaratives was measured in the same manner as in Benjamin and Schwanenflugel (2010) and in the same manner as described in Study 1. Henceforth, averaged sentence-final pitch change measurement from the grade level text is referred to as "level 3 sentence-final pitch," and from the above grade level text as "level 4 sentence-final pitch."

As discussed at length in Study 1, children's overall pitch variation within sentences, or pitch contour, was determined by measuring the mean pitch at the vocalic nucleus of each word within a sentence. All the measured pitches within a sentence were then averaged and a standard deviation was calculated. Standard deviations across sentences were averaged resulting in a mean pitch SD measure for each child. Children with higher standard deviations will likely be more fluent in general than children with lower standard deviations, since a higher standard deviation indicates greater pitch variation throughout a reading, while a low standard deviation may indicate "monotone" reading or word-by-word reading. Henceforth, this index of intonation contour from the grade level text will be referred to as “level 3 pitch SD,” and from the above grade level text as “level 4 pitch SD.”

Modifications to the scale. While Study 1 was designed as a development study, Study 2 was considered a validation study of the scale. In Study 2a I made a number of minor modifications to the scale based on expert feedback and confusions that arose regarding the use of the scale. Particular suggestions by expert raters were detailed in the *rater feedback* section of Study 1 results. Based on these suggestions and my own observations, I made changes to the expression dimension of the scale (both expressive intonation and natural pausing) as well as some changes to the instructions.

Raters had indicated that rater burden was adequate, averaging roughly four minutes per rating with approximately two of those minutes begin dedicated to the expression dimension of the scale. I modified scale instructions by including an introductory section explaining the purpose and nature of the scale. Finally, I also included instructions to raters to consider repetitions and hesitations as well as absolute pauses when rating children’s reading on the *natural pausing* dimension of the scale. The rationale behind this instruction was that it would better align with the method of measurement that had been used when measuring children’s pausing spectrographically for Study 2.

Changes made to the actual performance descriptions of the scale followed the recommendations of raters in Study 1, but were minor. First, the dimension title *expressive intonation* was changed to *appropriate intonation* for clarity and to discourage raters from equating expressive reading with dramatic reading. The goal was that children would read with appropriate expression, so the title of the dimension was changed to better reflect that goal. Some of the phrasing in the level 4 performance description of the appropriate intonation dimension was modified simply to avoid repetitiveness and made the description shorter. No substantive changes were made to this description, however. Finally, changes were made to the levels 2 and 4 performance descriptions of the natural pausing dimension. Line graphs from Benjamin and Schwanenflugel (2010) revealed that even the best readers would make an occasional misstep while reading aloud, so the level 4 performance description should reflect this. Good comprehension can take place even without perfect prosody. I also changed the level 2 performance description of the natural pausing dimension to reflect the possibility that readers may stumble significantly within sentences while reading, but may not necessarily mark sentence breaks with significant pausing. The critical feature of this performance level description should be frequent pausing, period. A struggling reader may or may not pause any more between sentences than he or she does within sentences. However, young readers at all skill levels do tend to have lengthier pauses between sentences than they do within sentences (Miller & Schwanenflugel, 2006), so I decided that the performance description should reflect both possibilities and revised it to state that readers may make significant pauses between sentences, as research has shown that struggling readers pause longer on average between sentences than skilled readers (Miller & Schwanenflugel, 2006). The revised scale used in Study 2a is found in Appendix F.

Expert rating procedures. Experts were sent a DVD with selected children's readings and copies of the revised rating scale, instructions, and passages. No formal training was conducted. However, raters had the opportunity to ask questions prior to beginning the ratings. Expert raters rated oral readings of the grade 3 story first, then rated oral readings of the grade 4 story. The order of oral readings was randomized. The raters independently conducted ratings in a quiet and isolated area free from distraction and background noise. Experts rated 60 children's oral readings of each passage plus five duplicate readings of each passage in order to obtain a consistency measure using the new scale. This resulted in each expert conducting a total of 130 ratings, 65 for each passage.

Study 2a Results

Prior to performing statistical analyses, data were analyzed for outliers using standard scores for all variables. Two possible outliers were found, with scores greater than three standard deviations from the mean scores. One participant's level 3 intersentential pause length was high for the subsample (*z*-score of 3.53) and a different participant's level 4 intersentential pause length was high for the subsample (*z*-score of 3.66), but after further examination it was found that these measurements had been conducted accurately and scores were retained. All standardized test scores and prosody measurements were examined for mean, range, standard deviation, skew, and kurtosis. All values were deemed initially acceptable based on the types of analyses being performed (see Table 3.1).

Data were also examined for effects of sex, race, and site on Comprehensive Oral Reading Fluency Scale total expression scores and comprehensive scores, standardized test scores, and spectrographic prosody measurements. One-way ANOVAs were performed and plots of means were examined. A Sex X Ratings ANOVA revealed an effect of sex for total expression scores with girls outperforming boys on the level 3 text, $F(1, 58) = 9.01$, $p = .004$. However, a Sex X Prosody ANOVA revealed an effect of sex for the Pitch SD variable for both the level 3 and level 4 passages, $F(1, 58) = 8.54$, $p =$

Table 3.1

Descriptive Statistics for Traditional Reading Assessments and Prosody Variables

	Minimum	Maximum	*M*	*Mdn*	*SD*
TOWRE	87	134	110.07	112.00	10.51
WIAT-RC	79	128	99.58	99.00	10.59
Level 3 QRI Fluency	66	170	116.14	116.56	28.44
Level 4 QRI Fluency	28	164	93.75	92.99	29.27
Level 3 QRI Comprehension	0	8	4.45	4.00	1.74
Level 4 QRI Comprehension	0	7	3.65	4.00	1.69
Level 3 Pause Ratio	.05	.52	.22	.21	.13
Level 4 Pause Ratio	.00	.80	.29	.30	.19
Level 3 Intersentential Pause Length	50	1453	555.70	540.17	254.28
Level 4 Intersentential Pause Length	129	1737	619.82	548.50	305.01
Level 3 Sentence-final Pitch	-37.40	97.27	24.82	19.14	26.93
Level 4 Sentence-final Pitch	-46.13	75.17	22.89	18.28	22.20
Level 3 Pitch SD	8.94	51.01	24.97	23.09	10.19
Level 4 Pitch SD	11.06	49.08	23.96	22.11	9.59

Note: *n* = 60. “Level 3” = level 3 passage; “level 4” = level 4 passage. TOWRE = Test of Word Reading Efficiency. WIAT-RC = Wechsler Individual Achievement Test, Reading Comprehension subtest. The normed mean for both the TOWRE and the WIAT-RC is 100 (*SD* = 15).

.005; *F*(1, 58) = 6.94, *p* = .011, respectively, with girls demonstrating greater pitch variation than boys. Thus, higher ratings can likely be explained by the girls’ more pronounced pitch variation. No other significant effects were found.

Interrater reliability. Interrater reliability was examined using two methods: 1) rater agreement percentages and 2) intraclass correlations. Intraclass correlation coefficients (ICCs; as opposed to *inter*class correlations, e.g., Pearson *r*) are generally obtained when comparisons of scores within participants on the same assessment are desirable (McGraw & Wong, 1996). Conventions for qualitatively describing the strength of ICCs are drawn from descriptions used for Kappas (Landis & Koch, 1977), such that an ICC of less than .40 is considered “poor,” between .40 and .59 is considered “moderate,” between .60 and .79 is considered “substantial,” and an ICC above .80 is “outstanding.” Henceforth, terminology used to describe ICCs will reflect these descriptors. In the present study, two raters measured WCPM for each of 60 participants and also rated each participant using the Comprehensive Oral Reading Fluency Scale. Based on the criteria set forth by Shrout and Fleiss (1979), for the present study in which two raters each rated the entire sample of participants, an ICC obtained through a Participant X Raters two-way random effects ANOVA is the appropriate method of analysis to use and is the method of ICC used throughout studies 2a and b. All intraclass correlation analyses conducted in this study use absolute agreement rather than simply consistency as the standard for comparisons and all ICCs reported are based on a single-measure rather than average-measure analysis.

Few consistent differences were found between text levels. Thus, for the sake of simplicity in understanding and comparing results across studies, interrater agreement percentages and ICCs reported in the text will reflect analyses in which the level 3 text and level 4 text ratings were analyzed together. Separate text analyses of interrater agreement percentages and ICCs can be found in Appendix G. Descriptive statistics of ratings, including the mean, median, mode, SD, and range of each rater are found in Table 3.2.

WCPM and the Rate and Accuracy dimension. Descriptive statistics for students’ WCPM scores on the reading selection assessed by raters can be found in Table 3.3. WCPM data resembled a Normal distribution. Percent agreement was used to

Table 3.2

Descriptive Statistics of Ratings for Each Rater

	Minimum	Maximum	*M*	*Mdn*	Mode	*SD*
			Level 3			
Rater 1 - Rate & Accuracy	2	8	5.50	6.00	4	1.94
Rater 2 - Rate & Accuracy	2	8	5.50	6.00	4	1.94
Rater 1 - Intonation	1	4	2.65	3.00	3	0.88
Rater 2 - Intonation	1	4	2.88	3.00	2	0.85
Rater 1 - Pausing	1	4	2.37	2.00	2	0.80
Rater 2 - Pausing	1	4	2.88	3.00	2	0.85
Rater 1 - Total Expression	2	8	5.02	5.00	5	1.51
Rater 2 - Total Expression	3	8	5.77	6.00	4	1.60
Rater 1 - Total Score	4	16	10.52	11.00	11	3.16
Rater 2 - Total Score	5	16	11.27	12.00	8	3.38
			Level 4			
Rater 1 - Rate & Accuracy	2	8	4.40	4.00	2	2.14
Rater 2 - Rate & Accuracy	2	8	4.37	4.00	2	2.07
Rater 1 - Intonation	1	4	2.10	2.00	2	0.90
Rater 2 - Intonation	1	4	2.65	3.00	2	0.90
Rater 1 - Pausing	1	4	1.95	2.00	2	0.85
Rater 2 - Pausing	1	4	2.38	2.00	2	0.90
Rater 1 - Total Expression	2	8	4.05	4.00	3	1.56
Rater 2 - Total Expression	2	8	5.03	5.00	4	1.72
Rater 1 - Total Score	4	16	8.45	8.00	5	3.40
Rater 2 - Total Score	4	16	9.40	9.00	8	3.66

Note: *n* = 60.

Table 3.3

Descriptive Statistics of WCPM as Assigned by Raters

	Minimum	Maximum	*M*	*SD*
L3 Rater 1 - WCPM	57	187	116.20	31.49
L3 Rater 2 - WCPM	56	182	115.30	31.32
L4 Rater 1 - WCPM	27	182	97.58	33.65
L4 Rater 2 - WCPM	26	182	96.58	33.38

Note: *n* = 60. L3 = Level 3; L4 = Level 4; WCPM = Words Correct Per Minute examine agreement among the ratings of the Rate and Accuracy dimension. Agreement among raters on the rate and accuracy dimension was high, as would be expected on a rating dimension that is based on the objective measurement of WCPM. Raters had 96% exact and 100% adjacent agreement on the rate and accuracy dimension. An intraclass correlation coefficient was obtained to determine reliability among students' rate and accuracy ratings among both raters. The ICC for rate and accuracy among raters was outstanding, ICC > .98, *F*(119, 119) = 103.09, *p* < .001.

Expression dimensions. Percent agreement analyses were conducted to determine the level of interrater agreement between the raters. Ratings for *Appropriate Intonation* and *Natural Pausing* each ranged from one to four. Due to the limited range of scores on the individual dimensions as well as the need for agreement on the ratings rather than simply consistency, the most significant analysis of these ratings is that of percent agreement. Agreement among raters on the *Appropriate Intonation* and *Natural Pausing* dimensions was low (57% and 51%, respectively) with generally greater agreement among ratings of intonation vs. pausing behavior. Adjacent agreement (95% for both dimensions) was high, however, but on a scale with scores ranging from one to four, such high adjacent agreement would be expected (cf. Daane et al., 2005; Pinnell et al, 1995). Intraclass correlations were conducted for appropriate intonation and natural pausing ratings as an additional measure of interrater reliability. Between the raters, the appropriate intonation ICC was substantial, ICC = .67, *F*(119, 119) = 6.41, *p* < .001. The natural pausing ICC was also substantial, ICC = .64, *F*(119, 119) = 6.37, *p* < .001. Thus, reliability was good and was similar across expression dimensions.

Total expression scores. Percent agreement was analyzed for total expression scores with the goal commonly being that adjacent agreement of 90% or higher among highly trained raters might be achieved. For raters without formal training, however, the 90% adjacent agreement may not be realistic. Adjacent agreement for total expression scores was 78% (exact agreement 31%). The ideal of 90% was not reached. Results were encouraging, however, from intraclass correlation analyses. An intraclass correlation was conducted for the Total Expression scores. Between the raters, the ICC was substantial, ICC = .73, $F(119, 119) = 10.48$, $p < .001$. Interrater reliability ability for total expression scores, then, is good but not excellent. Additional revisions to the scale may be necessary to improve interrater reliability without having to attach a training component to the scale.

Comprehensive oral reading fluency scores. Percent agreement was analyzed for total scale ratings with the goal being at least 90 percent agreement of +/- 2 points (cf. Rasinski et al., 2009). Agreement of 95% (exact = 30%; adjacent = 77%) was reached. High interrater agreement was achieved, indicating that total fluency may be reliably measured using the Comprehensive Oral Reading Fluency Scale. An intraclass correlation between the raters was conducted for the total scale scores. Between the raters, the ICC was outstanding, ICC = .93, $F(119, 119) = 48.48$, $p < .001$. While the high agreement among rate and accuracy ratings obviously contributed substantially to this high coefficient, it is evident from both the percent agreement analysis and the ICC that participants' comprehensive oral reading fluency scores can be interpreted as reliable.

Relationship between expression ratings and prosody measurements. Correlational relationships between ratings and prosody variables were examined to investigate the scale’s potential as an assessment that can be used to roughly gauge children’s prosody. To simplify interpretation within the study and comparisons across studies, the two raters’ ratings were averaged for each dimension; i.e., if rater 1 gave a child a rating of 3 for appropriate intonation and rater 2 gave the child a rating of 4, then a score of 3.5 is used to correlate the appropriate intonation dimension with prosody variables. Additionally, analyses reported here were conducted using both passages; thus, all intonation ratings (for both the level 3 and the level 4 texts) were correlated with all pitch SD measurements (both the level 3 and level 4 passage measurements). This method is used for all in-text results reported in this section. Correlations of individual raters’ ratings with corresponding prosody variables can be found in Appendix H.

Appropriate intonation. Ratings were averaged to examine the correlational relationships between the ratings and related prosody measurements: sentence-final pitch change, pause ratio, and pitch SD. These were all variables which had loaded onto the same factor, titled appropriate intonation, based on the factor analyses performed in phase 1 of Study 1. Using a linear Pearson correlation, mean appropriate intonation ratings correlated moderately to strongly with all prosody variables: sentence-final pitch (r = .42, p < .001), pause ratio (r = -.71, p < .001), and pitch SD (r = .59, p < .001).

Natural pausing. Ratings were averaged to examine the correlational relationships between the ratings and related prosody measurements: pause ratio and intersentential pause length. Pause ratio was used by Benjamin and Schwanenflugel (2010) and was found be a significant predictor of children’s automaticity and reading comprehension. The second variable, intersentential pause length, is a pause variable which was used in other studies (Miller & Schwanenflugel, 2006; Schwanenflugel et al.,

2004) and also found to be a useful predictor of reading skill. Using a linear Pearson correlation, mean natural pausing ratings correlated strongly with pause ratio (r = -.78, p < .001) but only moderately with intersentential pause length (r = -.31, p = .001).

Total expression. As with appropriate intonation and natural pausing ratings, total expression ratings were averaged across raters to examine the correlational relationships between the ratings and all prosody measurements. Using Pearson correlations, mean total expression scores were correlated with prosody variables: sentence-final pitch (r = .37, p < .001), pitch SD (r = .48, p < .001), pause ratio (r = -.78, p < .001), and intersentential pause length (r = -.25, p = .006). The correlation with intersentential pause length was relatively low, indicating that the scale's performance descriptions may need adjustment to better reflect this prosodic feature.

Relationship between total rating scores and traditional reading assessments. Correlational relationships between ratings and traditional reading assessments were examined to investigate the scale's potential as an assessment that can be used to roughly gauge children's prosody. To simplify interpretation within the study and comparisons across studies, the two raters' ratings were averaged for each dimension; i.e., if rater 1 gave a child a score of 5 for total expression and rater 2 gave the child a score of 7, then a score of 6 is used to correlate the total expression score with the other assessments. Additionally, analyses reported here were conducted using both passages; thus, all total expression ratings (for both the level 3 and the level 4 texts) were correlated with all QRI fluency scores (both the level 3 and level 4 passage measurements). This method is used for all in-text results reported in this section. Correlations separated by passage can be found in Appendix H.

Total expression scores. Correlational relationships were examined between total expression scores and traditional tests of text reading fluency (QRI fluency), passage comprehension (QRI comprehension), word reading fluency (TOWRE), and general reading comprehension (WIAT-RC). Correlations are reported in Table 3.4. Results suggest a strong relationship between children's prosody ratings and automaticity, and a moderate relationship between children's prosody ratings and general reading comprehension. Additionally, the correlation between total expression and WIAT-RC scores was equivalent to correlations between WIAT-RC scores and all other traditional reading assessments (see Table 3.4), providing evidence that simple ratings of reading expression can be used as a rough gauge of children's reading skill.

Table 3.4

Correlations Between Scale Scores and Traditional Reading Assessments

	Total Expression	Total Scale Score	QRI Fluency	QRI Compre-hension	TOWRE
Total Scale Score	.942**				
QRI Fluency	.851**	.924**			
QRI Comprehension	.205*	.239**	.263**		
TOWRE	.565**	.606**	.662**	.054	
WIAT-RC	.480**	.463**	.435**	.421**	.431**

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

As the QRI comprehension assessment is an informal assessment of comprehension that has not been normed or tested for its psychometric properties, it is difficult to interpret the low correlation between prosody ratings and QRI comprehension scores. Correlations among the traditional assessments, however, revealed a weak relationship between the QRI comprehension scores and other tests of reading ability.

Low correlations between passage comprehension and other tests of reading skill are not unusual in elementary school children (e.g, Valencia et al., 2010), so this result was not surprising. Because of the questionable validity of the QRI comprehension scores, results should not figure heavily into arguments for or against the validity of the Comprehensive Oral Reading Fluency Scale.

Comprehensive oral reading fluency scores. Total scale scores were averaged across raters to examine the correlational relationships between scale scores and traditional tests of text reading fluency (QRI fluency), passage comprehension (QRI comprehension), word reading fluency (TOWRE), and general reading comprehension (WIAT-RC). Correlations are reported in Table 3.4. Results suggest a strong relationship between children's overall performance on the Comprehensive Oral Reading Fluency Scale and their automaticity in reading both connected texts and word lists. This relationship is expected since a 50% weight in the comprehensive scale score is given to the rate and accuracy rating. Results suggest a moderate relationship between children's overall scale scores and their general reading comprehension, similar to reading comprehension's relationship with other traditional reading assessments. Issues with the QRI comprehension assessment are discussed in the section above.

Study 2a Rater Feedback

Following Study 2a, expert raters provided open-ended responses to requests for feedback about rater burden, the performance descriptions of the scale in general, their ability to distinguish between performance levels when rating a child's oral reading, clarity of language, and instructions provided for the scale. By counting the time it took them to rate a subsample of oral readings (approximately 20% of the sample), raters were able to provide an estimate of rater burden. Raters indicated that it took them an average of four minutes per rating to complete the 130 ratings and WCPM analyses for Study 2a, approximately 8.7 hours total. However, teachers already evaluate fluency by calculating WCPM, so really the scale is just adding the burden of the expressiveness ratings. The experts estimated that the additional burden provided by the expressiveness ratings was 2 minutes per child. They indicated that this was a reasonable amount of time for researchers who have recordings of children's oral readings and assistants to help with the workload.

Expert raters also suggested that the performance descriptions provide more concrete guidance to help determine, for example, how much pausing a child can do and still earn a "3" on the *Natural Pausing* dimension. One rater suggested that operational definitions of pausing and intonation be provided along with a brief introduction to the scale. This would allow raters to know what can count as a pause and what exactly they should be listening for regarding intonation.

Raters also made suggestions regarding the format of the scale. One rater suggested that some basic rater instruction be provided on the same page as the scale; such instruction should include operational definitions of intonation and pausing, so that raters who may be less familiar with measuring prosody might still be able to use the scale effectively. A rater also suggested adding an additional column between the appropriate intonation and natural pausing dimensions, so that raters can circle separate ratings for each dimension, making final calculations more straightforward. Finally, a rater suggested using some shading to make heading boxes more distinct.

Study 2b Method

Participants, general assessments and procedures, reading prosody assessment and procedures, and measurement of prosodic features were the same across studies 2a and 2b. Participants for Study 2b comprised the subsample of children (*n* = 60, from the larger *N* = 120) which were not included in the subsample for Study 2a.

Modifications to the scale. Based on feedback from the expert raters and analyses conducted using the ratings, the scale was revised accordingly. I made significant changes to the performance descriptions of the scale by examining students' prosodic measurements on line graphs along with their performance on standardized tests. Performance descriptions for *Natural Pausing* were made more concrete by dividing children into quartiles based on their standardized test performance and looking at the within- and between-sentence pausing that characterized students within each quartile. Each of the four performance levels, then, match up with the characteristics of students' pausing within each quartile.

It was more difficult, but not impossible, to make the performance descriptions for *Appropriate Intonation* more concrete. Again, students were examined in quartiles based on their standardized test performance, and characteristics of their sentence-final pitch and their pitch SD were used to modify performance descriptions. For example, line graphs revealed that each quartile's mean pitch SD lined up consistently with comprehension and word reading skills. That is, students in the lowest quartile had the lowest mean pitch SD and so on. Thus, more skilled readers have more pitch variation within sentences than less skilled readers. Additionally, while the most skilled readers consistently dropped their pitch noticeably at the end of declarative sentences and the least skilled readers tended to be flat, there was no real reliable difference between students in the second and third quartiles. Thus, performance descriptions at levels 2 and 3 were revised to focus more on appropriate overall intonation.

Expert rating procedures. Expert rating procedures differed in a few minor areas based on rater feedback from Study 2a. The amended scale can be found in Appendix I. Experts were sent a DVD with selected children's readings, a copy of the new rating scale, instructions for the rating scale, and a list detailing the order in which participants were to be listened to and assessed. No formal training was conducted, but raters were asked to conduct two test ratings on participants not included in Study 2b. Raters reported their ratings and were found to have perfect agreement, so no further test ratings were conducted. Both experts rated participants in the same order: the order of participants was randomized and alternated between stories. For example, raters would rate participant 1047 on story 1 followed by participant 2786 on story 2 and so on until all participants had been rated on both stories. The raters independently conducted ratings in a quiet and isolated area free from distraction and background noise. Experts rated 60 children's oral readings of each passage plus five duplicate readings of each passage in order to obtain a consistency measure using the new scale. This resulted in each expert conducting a total of 130 ratings, 65 for each passage.

Study 2b Results

Prior to performing statistical analyses, data were analyzed for outliers using standard scores for all variables. Five possible outliers were found, with scores greater than three standard deviations from the mean scores. One participant's TOWRE score was low for the subsample (*z*-score of -3.00) and this child's level 3 pause ratio was high (*z*-score of 3.64). Another participant's level 4 QRI fluency score was high for the subsample (*z*-score of 3.76). One participant's level 4 intersentential pause length was high (*z*-score of 3.45). Finally, one participant's level 3 sentence-final pitch change was high (*z*-score of 3.41). While these scores all were greater than three standard deviations from the mean, none of them appeared to be implausible. Further examination revealed that these measurements had been conducted accurately and scores were retained. All standardized test scores and prosody measurements were examined for mean, range, SD, skew, and kurtosis. All values were deemed initially acceptable based on the types of analyses being performed. Descriptive statistics can be found in Table 3.5.

Data were also examined for effects of sex, race, and site on Comprehensive Oral Reading Fluency Scale total expression scores and comprehensive scores, standardized test scores, and spectrographic prosody measurements. One-way ANOVAs were performed and plots of means were examined. A Sex X Ratings ANOVA revealed an effect of sex for total expression scores with girls outperforming boys on the level 3 text, *F*(1, 58) = 9.51, *p* = .003. Girls also outperformed boys on comprehensive scale scores for the level 3 passage, *F*(1, 58) = 8.39, *p* = .005. However, a Sex X Prosody ANOVA revealed an effect of sex on the sentence-final pitch change variable for the level 3 passage, *F*(1, 56) = 5.06, *p* = .028, with girls demonstrating greater sentence-final pitch declinations than boys. Thus, higher ratings may be explained by the girls' more appropriate pitch declinations at the ends of sentences. No other significant effects were found.

Interrater reliability. All methods relating to the analysis of interrater reliability were the same as those used in Study 2a. As with the previous study, few consistent

Table 3.5

Descriptive Statistics for Traditional Reading Assessments and Prosody Variables

	Minimum	Maximum	*M*	*Mdn*	*SD*
TOWRE	71	127	105.85	108.00	11.60
WIAT-RC	79	135	99.88	99.00	10.76
Level 3 QRI Fluency	39	196	110.27	113.01	31.91
Level 4 QRI Fluency	25	205	88.52	91.25	30.95
Level 3 QRI Comprehension	1	8	4.68	5.00	1.78
Level 4 QRI Comprehension	0	6	3.15	3.00	1.72
Level 3 Pause Ratio	.00	.76	.23	.19	.15
Level 4 Pause Ratio	.00	.85	.31	.30	.20
Level 3 Intersentential Pause Length	49	1856	663.94	572.17	409.58
Level 4 Intersentential Pause Length	83	1900	644.81	569.00	363.62
Level 3 Sentence-final Pitch	-50.76	120.10	21.02	16.55	29.05
Level 4 Sentence-final Pitch	-59.30	96.90	20.96	17.50	25.66
Level 3 Pitch SD	10.64	47.94	23.62	21.69	9.12
Level 4 Pitch SD	5.82	45.63	22.36	20.69	9.24

Note: *n* = 60. “Level 3” = level 3 passage; “level 4” = level 4 passage. TOWRE = Test of Word Reading Efficiency. WIAT-RC = Wechsler Individual Achievement Test, Reading Comprehension subtest. The normed mean for both the TOWRE and the WIAT-RC is 100 (*SD* = 15).

differences were found between text levels. Thus, for the sake of simplicity in understanding and comparing results across studies, interrater agreement percentages and ICCs reported in the text will reflect analyses in which the level 3 text and level 4 text ratings were analyzed together. Separate text analyses of interrater agreement percentages and ICCs can be found in Appendix J. Descriptive statistics of ratings, including the mean, median, mode, *SD*, and range of each rater are found in Table 3.6.

WCPM and the Rate and Accuracy dimension. Descriptive statistics for students’ WCPM scores on the reading selection assessed by raters can be found in Table 3.7. WCPM data resembled a Normal distribution. Percent agreement was used to

Table 3.6

Descriptive Statistics of Ratings for Each Rater

	Minimum	Maximum	*M*	*Mdn*	Mode	*SD*
			Level 3			
Rater 1 - Rate & Accuracy	2	8	5.17	6.00	4	1.99
Rater 2 - Rate & Accuracy	2	8	5.13	5.00	4	2.00
Rater 1 - Intonation	1	4	2.63	3.00	2	0.96
Rater 2 - Intonation	1	4	2.53	2.50	2	0.85
Rater 1 - Pausing	1	4	2.60	3.00	2[a]	0.85
Rater 2 - Pausing	1	4	2.70	3.00	3	1.01
Rater 1 - Total Expression	2	8	5.23	5.00	5	1.69
Rater 2 - Total Expression	2	8	5.23	5.00	5[a]	1.74
Rater 1 - Total Score	4	16	10.40	10.00	8	3.43
Rater 2 - Total Score	4	16	10.37	10.00	12	3.66
			Level 4			
Rater 1 - Rate & Accuracy	2	8	4.03	4.00	4	2.03
Rater 2 - Rate & Accuracy	2	8	3.97	4.00	2	2.00
Rater 1 - Intonation	1	4	2.32	2.00	3	0.98
Rater 2 - Intonation	1	4	2.02	2.00	1[a]	0.93
Rater 1 - Pausing	1	4	2.17	2.00	2	0.91
Rater 2 - Pausing	1	4	2.07	2.00	1	0.95
Rater 1 - Total Expression	2	8	4.48	5.00	5	1.74
Rater 2 - Total Expression	2	8	4.08	4.00	2	1.78
Rater 1 - Total Score	4	16	8.53	9.00	9	3.63
Rater 2 - Total Score	4	16	8.05	8.50	4	3.68

Note: *n* = 60.
a. Multiple modes exist. The smallest value is shown.

Table 3.7

Descriptive Statistics of WCPM as Assigned by Raters

	Minimum	Maximum	*M*	*SD*
L3 Rater 1 - WCPM	43	202	111.18	31.99
L3 Rater 2 - WCPM	42	178	110.03	30.59
L4 Rater 1 - WCPM	21	209	91.12	35.78
L4 Rater 2 - WCPM	23	170	89.45	33.44

Note: *n* = 60. L3 = Level 3; L4 = Level 4; WCPM = Words Correct Per Minute examine agreement among the ratings of the Rate and Accuracy dimension. Agreement among raters on the rate and accuracy dimension was high, as would be expected on a rating dimension that is based on the objective measurement of WCPM. Raters had 98% exact and 100% adjacent agreement on the rate and accuracy dimension. An intraclass correlation coefficient was obtained to determine reliability among students' rate and accuracy ratings among both raters for both stories. The ICC for rate and accuracy among raters was outstanding, ICC > .99, *F*(119, 119) = 174.81, *p* < .001.

Expression dimensions. Percent agreement analyses were conducted to determine the level of interrater agreement between the raters. Ratings for Appropriate Intonation and Natural Pausing each ranged from one to four. Due to the limited range of scores on the individual dimensions as well as the need for agreement on the ratings rather than simply consistency, the most significant analysis of these ratings is that of percent agreement. Exact agreement among raters on the Appropriate Intonation and Natural Pausing dimensions was low (57% and 60%, respectively) though somewhat improved since Study 2a. Adjacent agreement (98% for both dimensions) was high, however, but on a scale with scores ranging from one to four, such high adjacent agreement would be expected (cf. Daane et al., 2005; Pinnell et al, 1995). Intraclass correlations were conducted for appropriate intonation and natural pausing ratings as an additional measure of interrater reliability. Between the raters, the appropriate intonation ICC was substantial, ICC = .74, *F*(119, 119) = 7.12, *p* < .001. The natural pausing ICC was also substantial, ICC = .76, *F*(119, 119) = 7.25, *p* < .001. Thus, reliability was good and was similar across expression dimensions.

Total expression scores. Percent agreement was analyzed for total expression scores with the goal commonly being that adjacent agreement of 90% or higher among highly trained raters might be achieved. For raters without formal training, however, the 90% adjacent agreement may not be realistic. Adjacent agreement for total expression scores was 81% (exact agreement 44%). The ideal of 90% was not reached but numbers improved from Study 2a. Results from intraclass correlation analyses were encouraging. An intraclass correlation was conducted for the Total Expression scores. Between the raters, the ICC was outstanding, ICC = .81, $F(119, 119) = 9.90$, $p < .001$. Interrater reliability ability for total expression scores, then, is very good, especially considering that no formal training took place and raters did not confer with one another prior to or during rating.

Comprehensive oral reading fluency scores. Percent agreement was analyzed for total scale ratings with the goal being at least 90 percent agreement of +/- 2 points (cf. Rasinski et al., 2009). Agreement of 97% (exact = 43%; adjacent = 78%) was reached. High interrater agreement was achieved, indicating that total fluency may be reliably measured using the Comprehensive Oral Reading Fluency Scale. An intraclass correlation between the raters was conducted for the total scale scores. Between the raters, the ICC was outstanding, ICC = .95, $F(119, 119) = 40.86$, $p < .001$. While the high agreement among rate and accuracy ratings obviously contributed substantially to this high coefficient, it is evident from both the percent agreement analysis and the ICC that participants' comprehensive oral reading fluency scores can be interpreted as reliable.

Relationship between expression ratings and prosody measurements. Correlational relationships between ratings and prosody variables were examined to investigate the scale's potential as an assessment that can be used to roughly gauge children's prosody. Procedures for conducting correlations were the same as those used in Study 2a. Correlations of individual raters' ratings with corresponding prosody variables can be found in Appendix K.

Appropriate intonation. Ratings were averaged to examine the correlational relationships between the ratings and related prosody measurements: sentence-final pitch change, pause ratio, and pitch SD. These were all variables which had loaded onto the same factor, titled appropriate intonation, based on the factor analyses performed in phase 1 of Study 1. Using a linear Pearson correlation, mean appropriate intonation ratings correlated moderately to strongly with all prosody variables: sentence-final pitch ($r = .32$, $p < .001$), pause ratio ($r = -.70$, $p < .001$), and pitch SD ($r = .54$, $p < .001$). Correlations were similar to those in Study 2a.

Natural pausing. Ratings were averaged to examine the correlational relationships between the ratings and related prosody measurements: pause ratio and intersentential pause length. Pause ratio was used by Benjamin and Schwanenflugel (2010) and was found be a significant predictor of children's automaticity and reading comprehension. The second variable, intersentential pause length, is a pause variable which was used in other studies (Miller & Schwanenflugel, 2006; Schwanenflugel et al., 2004) and also found to be a useful predictor of reading skill. Both of these variables had loaded onto the same factor, titled natural pausing, based on the factor analysis performance in phase 1 of Study 1. Using a linear Pearson correlation, mean natural pausing ratings correlated strongly with pause ratio ($r = -.76$, $p < .001$) and moderately with intersentential pause length (*r* = -.44, *p* = .001). Results were higher than those in Study 2a, but differences were not statistically significant.

Total expression. As with appropriate intonation and natural pausing ratings, total expression ratings were averaged across raters to examine the correlational relationships between the ratings and all prosody measurements. Using Pearson correlations, mean total expression scores were correlated with prosody variables: sentence-final pitch (*r* = .24, *p* = .010), pitch SD (*r* = .44, *p* < .001), pause ratio (*r* = -.77, *p* < .001), and intersentential pause length (*r* = -.40, *p* < .001). The correlation with intersentential pause length improved upon Study 2a, but differences were not statistically significant. Increased correlations with intersentential pause length were likely a result of modifications made to the scale between studies 2a and b in which raters reported substantial improvement to the natural pausing dimension.

Relationship between total scores and traditional reading assessments. Correlational relationships between ratings and traditional reading assessments were examined to investigate the scale's potential as an assessment that can be used to roughly gauge children's prosody. All methods for conducting correlations were the same as those used in Study 2a. Correlations separated by passage can be found in Appendix K.

Total expression scores. Consistent with the above analyses, total expression scores were averaged across raters to examine the correlational relationships between the scores and traditional tests of text reading fluency (QRI fluency), passage comprehension (QRI comprehension), word reading fluency (TOWRE), and general reading comprehension (WIAT-RC). Correlations are reported in Table 3.8. Results suggest a strong relationship between children's prosody ratings and automaticity, and a moderate relationship between children's prosody ratings and both general and passage-specific reading comprehension. Additionally, the correlation between total expression and WIAT-RC scores was equivalent to correlations between WIAT-RC scores and all other traditional reading assessments (see Table 3.8), providing evidence that simple ratings of reading expression can be used as a rough gauge of children's reading skill. Issues with the QRI comprehension assessment as a valid measure of reading comprehension are discussed in Study 2a.

Table 3.8

Correlations Between Total Scale Scores and Traditional Reading Assessments

	Total Expression	Total Scale Score	QRI Fluency	QRI Comprehension	TOWRE
Total Scale Score	.968**				
QRI Fluency	.853**	.902**			
QRI Comprehension	.329**	.313**	.322**		
TOWRE	.744**	.751**	.776**	.072	
WIAT-RC	.463**	.456**	.438**	.476**	.347**

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Comprehensive oral reading fluency scores. Total scale scores were averaged across raters to examine the correlational relationships between scale scores and traditional tests of text reading fluency (QRI fluency), passage comprehension (QRI comprehension), word reading fluency (TOWRE), and general reading comprehension (WIAT-RC). Correlations are reported in Table 3.8. Results suggest a strong relationship between children's overall performance on the Comprehensive Oral Reading Fluency Scale and their automaticity in reading both connected texts and word lists. This relationship is expected since a 50% weight in the comprehensive scale score is given to the rate and accuracy rating. Results suggest a moderate relationship between children's overall scale scores and their general reading comprehension and passage-specific reading comprehension, similar to reading comprehension's relationship with other traditional reading assessments.

Chapter 4: Discussion

The present studies were conducted to accomplish the following broad goals: 1) to design a scale for measuring the complex construct of oral reading fluency, 2) to test the scale, and 3) to evaluate the scale based on evidence that could be used to make a rationale argument for or against the scale's usefulness. Study 1, phases 1 and 2, were designed to develop the initial scale and to conduct a pilot test of the scale using data from a prior study (Benjamin & Schwanenflugel, 2010). Studies 2a and 2b were designed to test and revise the scale using new data. In Chapter 1 I discussed the validation framework (Kane, 1992; 2006) that would be used to guide and interpret results of these studies and laid out the inferences that should be made when interpreting scores from the Comprehensive Oral Reading Fluency Scale. The following discussion presents several inferences upon which these studies were designed and validation arguments to consider when using the Comprehensive Oral Reading Fluency Scale. Finally, I present limitations of the current validation study.

Inferences and Validation

In Chapter 1 I listed several statements which comprise the inferences for the validation framework for the scale. That is, these statements specify the inferences that can be made when interpreting scale scores and are based on Kane's (1992; 2006) validation framework. Responses to these statements detail the evidence forming the validation argument for the scale. These statements will be considered in their original sequence.

Development based on spectrographic measures. Expressiveness components of an oral reading fluency scale should be grounded in the prosodic structure of children's oral reading, which can be measured spectrographically. Numerous studies have now revealed the usefulness of spectrographic research in understanding prosody's role in children's reading (e.g., Benjamin & Schwanenflugel, 2010; Dowhower, 1987; Miller & Schwanenflugel, 2006; 2008; Schwanenflugel et al., 2004). However, current scales designed to measure children's expression and fluency in general have not been developed or tested using spectrographic data. Thus, the key element that sets the current scale apart from the rest is its spectrographically-grounded development. As illustrated in phase 1 of Study 1, the dimensionality and the content of the current scale were developed from spectrographic measures of children's reading prosody. Factor analysis revealed the dimensional structure of children's prosody, and graphs of prosodic behavior in children at different levels of reading skill guided the writing of the scale's performance descriptions. Revisions to the scale (as discussed in studies 2a and b) were also based on further specification of the scale using children's prosodic measurements. Based on these facts, and the evidence demonstrating the relationship between ratings and prosodic measures discussed below, it is reasonable to contend that the Comprehensive Oral Reading Fluency Scale is grounded in the spectrographically measured prosodic structure of children's oral reading.

Development based on theory and research. While the specific format, dimensionality, and performance descriptions of the scale's expression component should adhere to empirically-based data regarding reading expression, a valid rating scale must also be consistent with current definitions, research, and theory in children's reading fluency and should be useful for its assessment. Though Kuhn et al. (2010) published the definition of fluency that guided the current research, they are not alone in advocating a multi-dimensional perspective of the reading fluency construct (Daane et al., 2005; Hudson et al., 2005; Hudson et al., 2009; Pinnell et al., 1995; Rasinski et al., 2009; Samuels, 2006). While automaticity is an important component of general reading skill (Logan, 1997), rate and accuracy measures alone cannot measure reading fluency as a construct combining "accuracy, automaticity, and oral reading prosody, which, taken together, facilitate the reader's construction of meaning" (Kuhn et al., 2010). Nonetheless, rate and accuracy comprise an important component of oral reading fluency, and because of the abundant research designing and using such measures, I was able to take advantage of current WCPM norms (Hasbrouck & Tindal, 2006) in the development of the current scale.

Prior studies and theory also informed the development of the expressiveness components of the scale, and performance descriptions are consistent with classic research and theory (Clay & Imlach, 1971; Dowhower, 1987; 1991) as well as new (Benjamin & Schwanenflugel, 2010; Fuchs et al., 2001; Hudson et al., 2009; Kuhn et al., 2010; Valencia et al., 2010). Thus, the Comprehensive Oral Reading Fluency Scale can be used to assess children's reading fluency without neglecting any facet of reading fluency as it is currently understood.

Interrater reliability. A valid scale must be a reliable scale, and experts in children's reading should be able to use the scale with good inter-rater reliability. Two methods of examining interrater reliability were used consistently in all studies: percent agreement and intraclass correlation. Interrater reliability for rate and accuracy ratings were, as expected, very high. Because they are based on objective measurements of WCPM, I will not discuss them further here. The ratings of interest for interrater reliability analyses are the ratings within the expression dimension: specifically the intonation and pausing ratings and the total expression ratings. Finally, interrater reliability of comprehensive oral reading fluency scores are useful particularly for comparison with other rating scales of oral reading fluency (e.g., Rasinski et al., 2009).

Across studies, interrater agreement percentages remained relatively stable, with some improvement in exact agreement among intonation, total expression, and comprehensive oral reading fluency scores (see Table 4.1). Only one expert served as a common rater in Studies 1, and 2a and b. Thus, improvements in inter-rater agreement are most likely due to improvements to the Comprehensive Oral Reading Fluency Scale itself.

Table 4.1

Percent Interrater Agreement and Intraclass Correlations Across Studies

	Exact agreement			Adjacent agreement			Intraclass correlations[c]		
	Study 1[a]	Study 2a	Study 2b	Study 1[a]	Study 2a	Study 2b	Study 1[a]	Study 2a	Study 2b
Appropriate intonation	48	57	57	95	95	98	.56	.67	.74
Natural pausing	62	51	60	99	95	98	.71	.64	.76
Total expression	36	31	44	83	78	81	.75	.73	.81
Comprehensive score	35	30	43	100[b]	95[b]	97[b]	.93	.93	.95

a. Study 1 agreement is the mean agreement among the three raters
b. adjacent agreement of +/- 2 points.
c. all correlations are significant at p < .001

It may be helpful to compare these rater agreement percentages with those reported for other fluency rating scales. Young et al. (1996) reported 94.4% rater agreement and a reliability coefficient of .85 (reported as r = .85) when using the Allington (1983) scale. However, the ratings which were counted in this agreement analysis were actually means of three ratings from each rater. Additionally, raters were highly trained and agreement was considered to exist when mean ratings were within one point of one another. Since the Allington scale is a 6-point scale, 94.4% agreement is not really much different than the adjacent agreement results for appropriate intonation and natural pausing in the present studies using untrained raters. A reliability coefficient reported as r = .85 implies that a Pearson correlation was conducted, a less stringent measure than the intraclass correlations conducted in the present studies and less appropriate for gauging reliability when absolute agreement among raters is desired. Thus, it is difficult to compare results from the present studies with those of Young et al., and the Allington scale and the Comprehensive Oral Reading Fluency Scale are dissimilar in format, so such comparisons may not be very informative.

The NAEP scale, however, is a 4-point scale that is similar in some ways to the appropriate intonation and natural pausing dimensions of the present scale. The first NAEP study to use the scale (Pinnell et al., 1995) reported .70 reliability, 58% exact agreement, and 98% adjacent agreement. These results are quite similar to the ICCs and agreement percentages reported in the present studies for appropriate intonation and natural pausing. A later NAEP study (Daane et al., 2005) reported improved interrater reliability and agreement (ICC = .82; 81% exact agreement; 100% adjacent agreement). While the exact agreement reported in this NAEP study is certainly higher than the exact agreement achieved in the present studies, ICCs between the NAEP study and the present Study 2b are comparable. Again, it should be noted that NAEP raters participated in extensive multi-day training sessions. Such extensive training is likely not practical in most circumstances in which a fluency rating scale might be used. Raters in the present studies did not receive any training, but relied on their expertise.

Most similar to the present scale is the Multidimensional Fluency Rubric published in Rasinski et al. (2009). Both scales have three dimensions and incorporate both automaticity and expression in some way. Ratings were conducted independently by two raters in both the Rasinski et al. study and the present studies 2a and b. However, if raters for the Rasinski et al. study disagreed by more than one point on any of the three dimensions of the scale, a third rater was brought in. For the present studies 2a and b, a third rater was not used for assistance in the event of such disagreement as the goal in this validity study was to get an idea of the level of agreement that might exist if the scale were to be used widely by other reading experts. Rasinski et al. reported interrater agreement (defined as +/- 2 points) for the 12-point scale as .857, though it is unclear whether this is a percentage or a reliability coefficient. This result, however, is similar to both the adjacent agreement and intraclass correlation reported for total expression in Study 2b (see Table 4.1). Using the same agreement standard as Rasinski et al., however (+/- 2 points), the agreements reported for comprehensive scores in the present studies as well as the ICCs are much higher than those for the Multidimensional Fluency Scale. From this comparison, it is evident that very good reliability was achieved without having to formally train raters. It is likely, though, that further training will improve agreement across dimensions.

Relationship between ratings and prosody measures. Scale ratings of children’s oral reading should correspond with spectrographic measures of children’s prosody. Since the initial development of the scale was based on a principle components analysis conducted in phase 1 of Study 1, and the development and further revisions to performance descriptions were led by examination of children's prosodic performance when reading aloud (see Benjamin & Schwanenflugel, 2010; Miller & Schwanenflugel, 2006), scale scores should correlate moderately to strongly with spectrographic measurements of prosodic features. Ungrammatical pause ratio was eliminated in Study 1 as a variable to be used for comparison as it performed poorly in comparisons with scale scores and, more importantly, did not correlate even moderately with other prosody variables. Evidence from Benjamin and Schwanenflugel (2010) showed the variable did not assist in predicting comprehension either.

Children's pausing within sentences (pause ratio), as expected, correlated quite strongly with all expression dimensions of the scale across all three studies. This is expected as within-sentence pausing has consistently corresponded strongly with measures of reading skill (e.g, Benjamin & Schwanenflugel, 2010; Clay & Imlach, 1971; Dowhower, 1987; Herman, 1985; Miller & Schwanenflugel, 2006; 2008; Schwanenflugel et al., 2004). Better readers simply pause less within sentences. However, better readers also have shorter pauses between sentences (e.g, Schwanenflugel et al., 2006). While pilot data in Study 1 showed a strong relationship between intersentential pause length and the scale's natural pausing dimension, that correlation was not as high when using new data in Study 2a. However, modifications to the scale were made between Studies 2a and b, and correlations between ratings and intersentential pause length improved in Study 2b.

Scale intonation and total expression ratings consistently correlated moderately to strongly with the spectrographic measure of children's pitch variation (pitch SD) across studies. Correlations between ratings and sentence-final pitch change, however, did decrease some across studies. This is likely due to feedback from raters between studies, who noted that most of the third grade readers seemed to drop their pitch at the ends of sentences—even poorer readers did this sometimes. Thus, raters believed that appropriate overall pitch intonation did a better job distinguishing between readers than did sentence-final pitch change. Because raters believed that a stronger focus on appropriate pitch variation within sentences might better distinguish good readers from struggling readers, more emphasis in performance descriptions was placed on pitch variation. The test of whether or not this shift in emphasis was appropriate should be an examination of the correspondence between ratings and other measures of children's reading skill in the following section.

It is impossible to compare results such as these with prior studies or other rating scales, as the Comprehensive Oral Reading Fluency Scale is the first scale of its kind to be developed using spectrographic measurements of reading prosody and tested against such measures. While other researchers have utilized their intuitions regarding prosody and fluency research in developing their scales (Pinnell et al., 1995; Rasinski et al., 2009), the development of the scales' dimensions did not include extraction of factors or principle components. Testing of these scales, also, did not include comparing ratings with actual prosody measurements. Thus, the current scale benefits from showing correspondence between actual reading prosody and expert ratings using the scales.

Relationship between ratings and traditional assessments of reading skill. Scale ratings of children's oral reading should correlate strongly with measures of children's reading rate and accuracy and should correlate moderately with measures of reading comprehension. For assessment scores to have a high degree of validity, they must measure what they actually purport to measure (Crocker & Algina, 1986). One way of demonstrating that an assessment does this is by examining its development and also having experts examine it for content validity. The Comprehensive Oral Reading Fluency Scale's appropriateness for measuring fluency and, specifically, expression has been demonstrated in the responses to the first two arguments presented above. Content validity was established as experts reviewed each revision of the scale providing feedback and approving of the scale as a useful measurement tool. However, reading fluency scores from the current scale must also demonstrate adequate correspondence with assessments which measure other reading skills known to correlate with reading fluency.

In Study 1 pre-existing data was used to develop and run an initial test on the Comprehensive Oral Reading Fluency Scale. Relationships between prosodic variables and standardized assessments were known based on results published in Benjamin and Schwanenflugel (2010), and relationships between total expression and comprehensive scores with standardized tests were high. Of greater interest here, however, is the correspondence between these scale scores and the assessments used in studies 2a and b. Correlations between scale scores and measures of fluency were high and remained stable or increased across the two studies. Correlations between scale scores and passage comprehension (QRI Comprehension) improved, and correlations between scale scores and general comprehension (WIAT-RC) remained stable and moderate.

The stability or even improvement in the relationships between scale scores and other traditional assessments of reading skill lends to an interpretation of the scale's validity in conjunction with the other validity evidence presented above. While the scale's creation was based largely on the premise that tests of automaticity alone are not sufficient for measuring reading fluency (Fuchs et al., 2001; Kuhn et al., 2010), automaticity still plays a critical role in children's reading (Logan, 1997). Thus, total expression scores should correlate highly with measures of reading rate and accuracy (i.e., QRI Fluency and TOWRE assessments). Fluency has traditionally had a moderate to strong relationship with general reading comprehension, and that relationship is evident in cross-study results.

Other scales of reading fluency have also been shown to correlate well with measures of automaticity and comprehension. For example, Rasinski (1985) reported good relationships between third graders' scores from a modified version of Allington's (1983) scale and reading rate, accuracy, and general comprehension. Young et al. (1996) used the Allington scale and reported correlations quite similar to those in the present studies. NAEP studies (Daane et al., 2005; Grigg, Daane, Jin, & Campbell, 2003; Mullis, Campbell, & Farstrup, 1993 Pinnell et al., 1995) also reported good moderate to high relationships between scale scores and traditional assessments of reading skill. Finally, Rasinski et al. (2009) reported a strong relationship between third graders' scores on the Multidimensional Fluency Rubric and reading comprehension.

It is apparent, then, that the relationship between scale scores and other reading tests alone cannot make an argument for the validity of one assessment over another; many factors must be considered in a validity argument (Kane, 1992; Messick 1989; 1995). Steps have been taken in the present studies to develop a psychometrically valid and practically useful assessment for measuring children's reading fluency as a whole construct. The advantage of the present scale over other existing scales is not necessarily that the Comprehensive Oral Reading Fluency Scale correlates more robustly with other fluency or comprehension assessments. Rather, the strength of this scale is that it has been built, from the ground up, on objective measurements of the construct it purports to measure. That is, if reading fluency is comprised of automaticity, accuracy, and prosody, then it makes sense to use objective measurements of reading rate, accuracy, and prosodic features to build a scale that can be used to measure the whole construct. Raters reported an additional burden of only about two minutes beyond what it would take to conduct a more traditional but less complete WCPM analysis of a student's reading. Since few researchers have the time and resources to measure prosody spectrographically, a scale which provides valid scores of children's reading expression and requires little to no training to achieve substantial reliability is certainly a useful tool to have.

Limitations and Future Research

While many steps were taken to ensure the development of a reliable and valid reading fluency scale, some issues will need to be addressed in future studies. First, only reading experts were used as raters in the present studies. This was done because only experts could provide the deep foundational knowledge necessary for grounding the scale in current understanding of the research on reading fluency. Because of their knowledge and experience, they were able to provide guidance regarding how to capture verbally the scale's specification of prosodic features related to fluency. They were able communicate conceptions of what constitutes a good fluent reading, both theoretically and practically. They were able to provide input during the earliest phases of development on scale utility and likely use. These are fundamental issues that novices would not be able to bring to this early validation effort. Thus, we could say that the current scale, not only was grounded in spectrographic data related to reading fluency, but also on fundamental knowledge of reading fluency as captured by the structure of expert insights on the topic.

Regardless, the ultimate goal for this scale is for classroom use. Because the scale has not been validated for use by teachers or reading specialists, it cannot yet be recommended for use by classroom teachers. Further research will need to determine whether and how much training is needed for classroom teachers to use the scale reliably and validly.

The scale was also only validated on second and third grade children, and the final version of the scale was only tested on a moderate sample size of third grade children. I can only recommend the scale for use, then, with second and third grade children. Future studies, however, can easily determine the usefulness of the scale in a range of grade levels (see Rasinski et al., 2009). While this limitation should be addressed, it is important to note that the NAEP scale was initially designed for and only tested on fourth grade children (Daane et al., 2005; Pinnell et al., 1995). Nonetheless, the scale is often recommended for use among multiple grade levels (McKenna & Stahl, 2009). Valencia et al. (2010), for example, used it successfully with grades ranging from second through sixth. Thus, it is reasonably likely that the Comprehensive Oral Reading Fluency Scale may translate well to other grade levels and at least its rate and accuracy dimension has been specifically designed to do so.

Finally, the current studies used a limited selection of texts. Indeed, only three were used. One of these was an extremely brief narrative. The other two were longer informational texts. On the other hand, the studies included texts at varying levels of complexity. Variants of the scale were shown to be valid using texts that were slightly below, at, and above children's actual grade level, suggesting that one can use it to evaluate the fluency of oral readings of texts of varying levels of complexity. Still, it is important to show that the current scale can be validly used with a wider variety of texts and genres such as informational, narratives, and even poetry. This would provide evidence for the scale's use as an informative assessment of children's reading ability and would also help determine whether the scale can be used reliably across multiple texts.

It is evident that prosodic reading is part of fluent reading. However, the mechanism by which prosody may assist readers in comprehending texts is not yet known. Since children may not be able to develop good prosody in their reading without also developing automaticity (Logan, 1997), the assessment of children's reading expression should not replace a balanced and multifaceted reading development and assessment program. One danger in increasing prosody's focus in assessment is that it may become overemphasized in reading instruction—just as some have argued that reading speed has become overemphasized. Because knowledge of prosody's role in reading is limited, and there are inherent difficulties in measuring prosody, a focus on reading expression should not outweigh considerations of reading for understanding.

References

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

Allington, R. L., and Brown, S. 1979. *FACT: A multimedia reading program.* Milwaukee, WI: Raintree Press.

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

Audacity Developer Team. (2008). Audacity (Version 1.2.6) [Computer software]. Available: audacity.sourceforge.net/download/

Audacity Developer Team. (2011). Audacity (Version 1.3.14) [Computer software]. Available: audacity.sourceforge.net/download/

Aulls, M.W. (1978). *Developmental and remedial reading in the middle grades.* Boston, MA: Allyn & Bacon.

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

Blumstein, S., & Goodglass, H. (1972). The perception of stress as a semantic cue in aphasia. *Journal of Speech & Hearing Research*, *15*(4), 800-806.

Boersma, P., & Weenink, D. (2008). Praat: Doing phonetics by computer (Version 5.0.38) [Computer software]. Current version available: www.fon.hum.uva.nl/praat/

Boersma, P., & Weenink, D. (2011). Praat: Doing phonetics by computer (Version 5.2.22) [Computer software]. Current version available: www.fon.hum.uva.nl/praat/

Bolaños, D., Cole, R.A., Ward, W. H., Tindal, G. A., & Schwanenflugel, P.J. (under review). Automatic assessment of expressive oral reading.

Chafe, W. (1988). Punctuation and the prosody of written language. *Written Communication, 5*(4), 395-426.

Chall, J. (1983). *Stages of reading development*. New York, NY: McGraw-Hill.

Christ, T., & Ardoin, S. (2009). Curriculum-Based Measurement of Oral Reading: Passage Equivalence and Probe-Set Development. *Journal of School Psychology*, *47*(1), 55-75.

Clark, R., Morrison, T., & Wilcox, B. (2009). Readers' theater: A process of developing fourth-graders' reading fluency. *Reading Psychology*, *30*(4), 359-385.

Clay, M. M., & Imlach, R. H. (1971). Juncture, pitch, and stress as reading behavior variables. *Journal of Verbal Learning & Verbal Behavior, 10*(2), 133-139.

Couper-Kuhlen, E. (1986). *An introduction to English prosody*. London: Edward Arnold.

Cowie, R., Douglas-Cowie, E., & Wichmann, A. (2002). Prosodic characteristics of skilled reading: Fluency and expressiveness in 8-10-year-old readers. *Language and Speech*, *45*(1), 47-82. doi:10.1177/00238309020450010301.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart, & Winston.

Cromer, W. (1970). The difference model: A new explanation for some reading difficulties. *Journal of Educational Psychology*, *61*(6, Pt.1), 471-483. doi:10.1037/h0030288.

Cronbach, L.J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, 40(2), 141-201.

Daane, M.C., Campbell, J.R., Grigg, W.S., Goodman, M.J., & Oranje, A. (2005). *Fourth-grade students reading aloud: NAEP 2002 special study of oral reading* (NCES 2006-469). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Deeney, T. A. (2010). One-minute fluency measures: Mixed messages in assessment and instruction. *Reading Teacher*, 63(6), 440-450.

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

Deno, S., & Marston, D. (2006). Curriculum-based measurement of oral reading: An indicator of growth in fluency. *What research has to say about fluency instruction* (pp. 179-203). Newark, DE US: International Reading Association.

Dowhower, S. L. (1987). Effects of repeated reading on second-grade transitional readers’ fluency and comprehension. *Reading Research Quarterly*, *22*, 389-406. doi: 10.2307/747699

Dowhower, S.L. (1991). Speaking of prosody: Fluency's unattended bedfellow. *Theory Into Practice*, *30*(3), 165–175. doi:10.1080/00405849109543497

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. doi:10.1037/h0057532

Frazier, L., Carlson, K., & Clifton, C. (2006). Prosodic phrasing is central to language comprehension. *Trends in Cognitive Sciences*, *10*(6), 244-249. doi:10.1016/j.tics.2006.04.002.

Frick, T., & Semmel, M. (1978). Observer agreement and reliabilities of classroom observational measures. *Review of Educational Research*, *48*(1), 157-184. doi:10.2307/1169913.

Fuchs, L., Fuchs, D., Hosp, M., & Jenkins, J. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, *5*(3), 239-56.

Fuchs, L., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *RASE: Remedial & Special Education*, *9*(2), 20-28. doi:10.1177/074193258800900206.

Good, R.H., III, & Kaminski, R.A. (2002). *Dynamic indicators of basic early literacy skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement. Retrieved March 9, 2011, from dibels.uoregon.edu

Grigg, W.S., Daane, M.C., Jin, Y., and Campbell, J.R. (2003). *The nation's report card: Reading 2002* (NCES 2003–521). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Gronlund, N.E., & Waugh, C.K. (2009). *Assessment of student achievement* (9th ed.). Boston, MA: Pearson.

Hasbrouck, J., & Tindal, G. A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *Reading Teacher*, *59*(7), 636-644.

Herman, P. A. (1985). The effect of repeated readings on reading rate, speech pauses, and word recognition accuracy. *Reading Research Quarterly*, *20*(5), 553-565. doi:10.2307/747942

Hudson, R. F., Isakson, C., Richman, T., Lane, H. B. & Arriaza-Allen, S. (2011). An examination of a small-group decoding intervention for struggling readers: Comparing accuracy and automaticity criteria. *Learning Disabilities Research & Practice, 26,* 15–27. doi: 10.1111/j.1540-5826.2010.00321.x

Hudson, R., Lane, H., & Pullen, P. (2005). Reading fluency assessment and instruction: What, why, and how?. *Reading Teacher*, *58*(8), 702-714.

Hudson, R., Pullen, P., Lane, H., & Torgesen, J. (2009). The complex nature of reading fluency: A multidimensional view. *Reading & Writing Quarterly*, *25*(1), 4-32.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527-535. doi:10.1037/0033-2909.112.3.527.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

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

Kleiman, G. M., Winograd, P. N. and Humphrey, M. H. (1979). *Prosody and children's parsing of sentences* (Tech. Rep. 123). Urbana, IL: University of Illinois, Center for the Study of Reading.

Koriat, A., Greenberg, S. N., & Kreiner, H. (2002). The extraction of structure during reading: Evidence from reading prosody. *Memory & Cognition, 30*(2), 270-280.

Kuhn, M. (2005). A comparative study of small group fluency instruction. *Reading Psychology*, *26*(2), 127-146. doi:10.1080/02702710590930492.

Kuhn, M., Schwanenflugel, P., & Meisinger, E. (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*, 3-21.

LaBerge, D., & Samuels, S. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*(2), 293-323. doi:10.1016/0010-0285(74)90015-2.

Lai, S.A., Benjamin, R.G, Schwanenflugel, P.J., & Kuhn, M.R. (in press). The longitudinal relationship between reading fluency and reading comprehension skills in second grade children. *Reading & Writing Quarterly*.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical Data. *Biometrics, 33*(1), 159-174.

Leslie, L., & Caldwell, J. (2011). *Qualitative reading inventory-5 (QRI-5).* Boston, MA: Pearson.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, 95(4), 492-527. doi:10.1037/0033-295X.95.4.492

Logan, G. D. (1997). Automaticity and reading: Perspectives from the instance theory of automatization. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, 13(2), 123-146. doi:10.1080/1057356970130203

McGraw, K. O., and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30-46.

McKenna, M., & Stahl, K. (2009). *Assessment for reading instruction (2nd ed.)*. New York, NY: Guilford Press.

Messick, S. (1989). Validity. In R. L. Linn, R. L. Linn (Eds.), Educational measurement (3rd ed.) (pp. 13-103). New York, NY England: Macmillan Publishing Co, Inc.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14*(4), 5-8. doi:10.1111/j.1745-3992.1995.tb00881.x.

Meyer, M., & Felton, R. (1999). Repeated reading to enhance fluency: Old approaches and new directions. *Annals of Dyslexia*, *49,* 283-306.

Miller, M.D., Linn, R.L., & Gronlund, N.E. (2009). *Measurement and assessment in teaching* (10th ed.). Boston, MA: Pearson.

Miller, J., & Schwanenflugel, P.J. (2006). Prosody of syntactically complex sentences in the oral reading of young children. *Journal of Educational Psychology*, *98*(4), 839–853. doi:10.1037/0022-0663.98.4.839

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

Mullis, I., Campbell, J.R., Farstrup, A.E. (1993). *NAEP 1992--Reading report card for the nation and the states: Data from the national and trial state assessments* (NAEP-23-ST06). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups* (NIH Publication No. 00-4754). Washington, DC: U.S. Government Printing Office.

Nikto, A.J. (2004). *Educational assessment of students* (4th ed.). Boston, MA: Pearson.

Nunez, L. (2009). *An analysis of the relationship of reading fluency, comprehension, and word recognition to student achievement* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses.

O'Shea, L., & Sindelar, P. (1983). The effects of segmenting written discourse on the reading comprehension of low- and high-performance readers. *Reading Research Quarterly*, *18*(4), 458-465. doi:10.2307/747380.

Pangrac, A. (2009). *An examination of DIBELS oral reading fluency scores and their relationship to comprehension as measured by the third grade Ohio reading achievement test* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses.

Perfetti, C. (1985). *Reading ability*. New York, NY: Oxford University Press.

Pinnell, G.S., Pikulski, J.J., Wixson, K.K., Campbell, J.R., Gough, P.B., & Beatty, A.S. (1995). *Listening to children read aloud: Data from NAEP's integrated reading performance record (IRPR) at grade 4* (NCES 95-726). Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Psychological Corporation. (1992). Wechsler individual achievement test [Assessment kit]. San Antonio, TX, Harcourt Brace Jovanovich.

Rasinski, T. V. (1985). A *study of factors involved in reader-text interactions that contribute to fluency in reading (phrasing, process, models, interactive)* (Doctoral dissertation). Retrieved from). ProQuest Dissertations and Theses.

Rasinski, T. (2004). *Assessing Reading Fluency*. Pacific Institute for Research and Evaluation.

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.

Ravid, D., & Mashraki, Y. E. (2007). Prosodic reading, reading comprehension and morphological skills in Hebrew-speaking fourth graders. *Journal of Research in Reading, 30*(2), 140-156.

Reutzel, D., Fawson, P., & Smith, J. (2008). Reconsidering silent sustained reading: An exploratory study of scaffolded silent reading. *Journal of Educational Research*, *102*(1), 37-50.

Reynolds, C.R., Livingston, R.B., & Willson, V. (2009). *Measurement and assessment in education* (2nd ed.). Boston, MA: Pearson.

Richardson, V., Anders, P., Tidwell, D., & Lloyd, C. (1991). The relationship between teachers' beliefs and practices in reading comprehension instruction. *American Educational Research Journal*, *28*(3), 559–586.

Roehrig, A., Petscher, Y., Nettles, S., Hudson, R., & Torgesen, J. (2008). Accuracy of the DIBELS oral reading fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology*, *46*(3), 343-366. doi:10.1016/j.jsp.2007.06.006.

Rumelhart, D. E. (1994). Toward an interactive model of reading. In R. B. Ruddell, M. Ruddell, H. Singer, R. B. Ruddell, M. Ruddell, H. Singer (Eds.) , *Theoretical models and processes of reading (4th ed.)* (pp. 864-894). Newark, DE US: International Reading Association.

Samuels, S. (1994). Toward a theory of automatic information processing in reading, revisited. In R. B. Ruddell, M. Ruddell, H. Singer, R. B. Ruddell, M. Ruddell, H. Singer (Eds.) , *Theoretical models and processes of reading (4th ed.)* (pp. 816-837). Newark, DE US: International Reading Association.

Samuels, S. (2006). Toward a model of reading fluency. *What research has to say about fluency instruction* (pp. 24-46). Newark, DE US: International Reading Association.

Samuels, S. (2007). The DIBELS tests: Is speed of barking at print what we mean by reading fluency?. *Reading Research Quarterly*, *42*(4), 563-566.

Sargent, S. (2002, April). Oral reading fluency: A predictor of reading proficiency in fifth-grade students?. Paper presented at the Annual Meeting of the International Reading Association, San Francisco, CA.

Schreiber, P. (1980). On the acquisition of reading fluency. *Journal of Reading Behavior*, *12*(3), 177-86.

Schreiber, P. A. (1987). Prosody and structure in children's syntactic processing. In R. Horowitz & S. J. Samuels (Eds.), *Comprehending oral and written language* (pp. 243–270). San Diego, CA: Academic Press.

Schreiber, P. A. (1991). Understanding prosody's role in reading acquisition. *Theory Into Practice*, *30*(3), 158-164.

Schwanenflugel, P.J., & Benjamin, R.G. (2012). Reading expressiveness: The neglected aspect of reading fluency. In Rasinski, Blachowicz, and Lems (Eds.), *Fluency Instruction (2nd ed.).* NY, NY: Gilford.

Schwanenflugel, P., Hamilton, A., Kuhn, M., Wisenbaker, J., & Stahl, S. (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. doi:10.1037/0022-0663.96.1.119

Share, D.L. (1999). Phonological recoding and orthographic learning: A direct test of the self-teaching hypothesis. *Journal of Experimental Child Psychology, 72,* 95-129.

Shelton, N., Altwerger, B., & Jordan, N. (2009). Does DIBELS put reading first?. *Literacy Research and Instruction*, 48(2), 137-148. doi:10.1080/19388070802226311

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420-428. doi:10.1037/0033-2909.86.2.420

Smith, C. L. (2004). Topic transitions and durational prosody in reading aloud: Production and modeling. *Speech Communication, 42*(3), 247-270.

Snow, D.P., & Coots, J.H. (1981). Sentence perception in listening and reading (Tech. Note 2-81/15). Los Alamitos, CA: Southwest Regional Laboratory.

Spache, G. (1953). A new readability formula for primary-grade reading materials. *The Elementary School Journal*, *53*(7), 410–413. doi:10.1086/458513

Spear-Swerling, L., & Sternberg, R. J. (1996). *Off track: When poor readers become 'learning disabled'*. Boulder, CO US: Westview Press.

Steinhauer, K. (2003). Electrophysiological correlates of prosody and punctuation. *Brain and Language*, 86(1), 142-164. doi:10.1016/S0093-934X(02)00542-4

Suchey, N. (2009). *Oral passage reading fluency as an indicator of reading comprehension*. PhD dissertation, Department of Special Education, The University of Utah, Salt Lake City, UT.

Surányi, Z., Csépe, V., Richardson, U., Honbolygó, F., Goswami, U., & Thomson, J. M. (2009). Sensitivity to rhythmic parameters in dyslexic children: A comparison of hungarian and english. *Reading and Writing, 22*(1), 41-56.

Swets, B., Desmet, T., Hambrick, D., & Ferreira, F. (2007). The role of working memory in syntactic ambiguity resolution: A psychometric approach. *Journal of Experimental Psychology: General*, *136*(1), 64-81. doi:10.1037/0096-3445.136.1.64.

Syrdal, A. K., Hirschberg, J., McGory, J., & Beckman, M. (2001). Automatic ToBI prediction and alignment to speed manual labeling of prosody. *Speech Communication, 33*(1-2), 135-151.

Tindal, G., & Marston, D. (1996). Technical adequacy of alternative reading measures as performance assessments. *Exceptionality*, *6*(4), 201-230. doi:10.1207/s15327035ex0604_1.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). Test of word reading efficiency [Assessment kit]. Austin, TX: Pro-Ed.

Valencia, S., Smith, A., Reece, A., Li, M., Wixson, K., & Newman, H. (2010). Oral Reading fluency assessment: Issues of construct, criterion, and consequential validity. *Reading Research Quarterly*, *45*(3), 270-291.

Wechsler, D. (2009). Wechsler individual achievement test (3rd ed.) [Assessment kit]. San Antonio, TX: Harcourt Assessment, Inc.

Wiederholt, J.L., & Bryant, B.R. (1992). GORT-3: Gray oral reading tests (3rd ed.) [Assessment kit]. Austin, TX: Pro-Ed.

Wiederholt, J.L., & Bryant, B.R. (2001). GORT-4: Gray oral reading tests (4th ed.) [Assessment kit]. Austin, TX: Pro-Ed.

Whalley, K., & Hansen, J. (2006). The role of prosodic sensitivity in children's reading development. *Journal of Research in Reading, 29*(3), 288-303.

Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Scientific Studies of Reading*, 5(3), 211-38.

Yıldız, M., Yıldırım, K., Ateş, S., Çetinkaya, Ç. (2008). An evaluation of the oral reading fluency of 4th graders with respect to prosodic characteristic. *International Journal of Human Sciences* [Online]. 6:1. Available: http://www.insanbilimleri.com.

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.

Young, A., Bowers, P., & MacKinnon, G. (1996). Effects of prosodic modeling and repeated reading on poor readers' fluency and comprehension. *Applied Psycholinguistics*, *17*(1), 59-84. doi:10.1017/S0142716400009462.

Zervas, P., Fakotakis, N., & Kokkinakis, G. (2008). Development and evaluation of a prosodic database for Greek speech synthesis and research. *Journal of Quantitative Linguistics*, *15*(2), 154-184.

Zutell, J., & Rasinski, T. (1991). Training teachers to attend to their students' oral reading fluency. *Theory into Practice*, *30*(3), 211-17.

Appendix A: Study 1 Factor Analysis Correlations

Table A1

Correlations among variables used for factor analysis

	Easy Sentence-final Pitch Change	Difficult Sentence-final Pitch Change	Easy Pause Ratio	Difficult Pause Ratio	Easy Ungrammatical Pause	Difficult Ungrammatical Pause	Easy Pitch SD
Difficult Sentence-final Pitch Change	.381**						
Easy Pause Ratio	-.316**	-.317**					
Difficult Pause Ratio	-.416**	-.438**	.842**				
Easy Ungrammatical Pause	-.039	-.067	.265*	.298**			
Difficult Ungrammatical Pause	-.325**	-.295**	.477**	.617**	.273**		
Easy Pitch SD	.716**	.631**	-.420**	-.557**	-.101	-.495**	
Difficult Pitch SD	.649**	.557**	-.410**	-.523**	-.214*	-.537**	.812**

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Appendix B: Study 1 First Draft of Comprehensive Oral Reading Fluency Scale

3rd grade: Spring

RATE & ACCURACY		EXPRESSION		
rating	Rate/Accuracy	rating	Intonation	Pause
8	137+ WCPM	4	Varies pitch appropriately throughout sentences to communicate meaning; makes appropriate and consistent end of sentence pitch changes	Within-sentence pauses are short and necessary to convey meaning. Between-sentence pauses are short.
6	107+ WCPM	3	Varies pitch appropriately most of the time; tends to drop pitch at the end of declarative sentences.	Has some longer pauses within and between sentences, but they don't significantly interrupt the flow of the text.
4	78+ WCPM	2	Intonation may often be flat or unpredictable; may not end sentences with appropriate pitch changes.	Frequent pausing within sentences and some inappropriate or lengthy pausing between sentences
2	<78 WCPM	1	Read with flat or unnatural intonation; does not mark the ends of sentences with appropriate pitch changes.	Reading is broken and with numerous pauses throughout.

Rate/Accuracy: ______ **Intonation: _______** **Pause: _______**

EXPRESSION: ______

TOTAL: ______

Instructions:

- Listen to students read a text which, on average, take one-minute or more using a text (at/above) the student's grade level.
- The numbers in the WCPM column should reflect the Hasbrouk & Tindal (2006) quartiles for the appropriate grade level and time of year.
- While listening to the student read, obtain WCPM by counting errors and subtracting from the WPM.
  - If the student reads for less than or longer than one minute, simply subtract the number of errors from the number of words read, then divide that number by the number of seconds the child read
  - The result is the WCPM. Based on the Hasbrouck & Tindal (2006) fluency norms, give them a score of 2 through 8.
- Also listen for intonation and pausing, and rate the child's reading on the scale of 1-4 for each category. Use the rating which most closely fits the child's reading in each category. Note, for example, that this scale can account for children who read quickly and accurately, with appropriate pausing, but with poor intonation.
- Note that while a total score is obtained, sub-scale scores are also important to keep track of as they diagnose where fluency problems may lie.

Appendix C: Study 1 Final Draft of Comprehensive Oral Reading Fluency Scale

2nd Grade: Spring

RATE & ACCURACY		EXPRESSION		
Rating	**Rate/Accuracy**	**Rating**	**Expressive Intonation**	**Natural Pausing**
8	117+ WCPM	4	Varies pitch appropriately throughout sentences to communicate meaning; makes appropriate and consistent end of sentence pitch changes; pause patterns work to convey expressiveness	Within-sentence pauses are short and necessary to convey meaning. Between-sentence pauses are short, but natural.
6	89+ WCPM	3	Varies pitch appropriately most of the time; tends to drop pitch at the end of declarative sentences. May try to correct the prosody to match the phrasing of the text after initially getting it wrong.	Has some longer pauses within and between sentences, but they only momentarily interrupt the flow of the text. Pauses seem to be used mainly to distinguish phrases and sentences.
4	61+ WCPM	2	Intonation often may be flat or not matching the meaning/phrasing of the text (though some attempt may be made); may often not end sentences with appropriate pitch changes.	Frequent pausing within sentences and some lengthy pausing between sentences
2	<61 WCPM	1	Reads with flat and unnatural intonation throughout; does not mark sentence boundaries with appropriate pitch changes.	Reading is broken and effortful with numerous pauses throughout.

Rate/Accuracy: ______ **Intonation:** _______ **Pause:** _______

EXPRESSION: ______ ***(Actual WCPM:** _______**)***

TOTAL RATING: ______

Instructions:

- Listen to students read a text which, on average, takes one-minute or more using a text (at/above) the student's grade level.
- The numbers in the WCPM column should reflect the Hasbrouk & Tindal (2006) quartiles for the appropriate grade level and time of year.
- While listening to the student read, obtain WCPM by counting errors and subtracting from the WPM.
    - Errors—use QRI guidelines for counting errors:
        - **inserting** a word should count as one error
        - **omitting** a word should count as one error
        - even if a student **self-corrects**, the original error is still counted
        - a **reversal** counts as one error (switching the order of two words or phrases in a text)
        - **skipping** a line should count as one error, and student should be directed back to read the line once it is evident that it has been skipped
        - a **mispronunciation** counts as one error each time the word is mispronounced, **except in the case of proper nouns**—if a student consistently uses the same wrong pronunciation of a proper noun, it only counts as a single error
        - after pausing for three seconds, the student should be directed to skip the word—this is then counted as one error
    - If the student reads for less than or longer than one minute, simply subtract the number of errors from the number of words read, then divide that number by the number of seconds the child read. Multiply by 60. The result is WCPM.

      Example: 75 words read – 3 errors = 72 correct words
      72 words correct / 42 seconds = 1.714
      1.714 x 60 = **103 WCPM**

    - Based on the Hasbrouck & Tindal (2006) fluency norms, give them a score of 2 through 8 on the rating scale.
- Also listen for intonation and pausing, and rate the child's reading on the scale of 1-4 for each category. Use the rating which most closely fits the child's reading in each category. Note, for example, that this scale can account for children who read quickly and accurately, with appropriate pausing, but with poor intonation.
- Note that while a total score is obtained, sub-scale scores are also important to keep track of as they diagnose where fluency problems may lie.

Appendix D: Study 1 Individual Rater Correlations

Table D1

Study 1 Pearson Correlations Between Individual Raters' Expression Ratings and Prosody Variables

Prosody Variables	Expressive Intonation			Natural Pausing			Total Expression		
	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3
Sentence-final pitch change	.48**	.52**	.49**	--	--	--	.51**	.50**	.49**
Pitch SD	.61**	.54**	.55**	--	--	--	.60**	.58**	.54**
Pause ratio	-.60**	-.56**	-.58**	-.76**	-.68**	-.77**	-.78**	-.73**	-.75**
Ungrammatical pause ratio	--	--	--	-.28*	-.27*	-.36**	-.25	-.26*	-.23
Intersentential pause length	--	--	--	-.51**	-.59**	-.60**	-.46**	-.58**	-.49**

* *p* < .05, ** *p* < .001

Table D2

Study 1 Pearson Correlations Between Individual Raters' Total Ratings and Standardized Tests

	GORT-rate	WIAT-RC	TOWRE	Rater 1 Total Expression	Rater 1 Total Scale Score	Rater 2 Total Expression	Rater 2 Total Scale Score	Rater 3 Total Expression
WIAT-RC	.746**							
TOWRE	.755**	.635**						
Rater 1 Total Expression	.815**	.726**	.678**					
Rater 1 Total Scale Score	.842**	.711**	.734**	.915**				
Rater 2 Total Expression	.807**	.634**	.713**	.751**	.810**			
Rater 2 Total Scale Score	.815**	.631**	.735**	.763**	.933**	.906**		
Rater 3 Total Expression	.780**	.673**	.687**	.832**	.836**	.801**	.807**	
Rater 3 Total Scale Score	.817**	.676**	.734**	.817**	.953**	.831**	.953**	.915**

**. Correlation is significant at the 0.01 level (2-tailed).

Appendix E: Studies 2a and b Procedures for Measuring Prosody

Passages to measure:
Story #1 – 3rd paragraph plus next two sentences
Story #2 – second sentence of the 3rd paragraph through the 5th sentence of the 4th paragraph. These segments were selected because, of the segments in the middle of the passage, these sentences came closest to averaging a minute for a random sampling of 30 students (M = 65 seconds) without having to cut the segment mid-sentence.

General rules:

- □ avoid measuring “cliffs” or sharp, unnatural, rises or falls in pitch as indicated by Pratt. In these cases the child is not actually adjusting their pitch as indicated by the software. Rather, the software has captured either some irrelevant phonetic information or some background noise.
- □ Sometimes when there is background noise or a child reads in “creaky voice”, the pitch will appear to be very low (i.e., typically below 150 Hz). In the case where a child is typically reading at a pitch of around 250 and then a word appears that *sounds* similar to the others but *measures* at a very low pitch, then this is a misreading by Praat and should not be counted as the actual pitch of the word. Sometimes adjusting the viewing window can result in Praat’s accurately capturing the pitch of the word.
- □ In some cases a child begins reading in “creaky voice” about halfway through a word. In this case, use judgment in determining if there is enough information to capture the general pitch of the word. If not, simply count it as missing data. If this takes place at the end of a sentence when measuring sentence-final pitch change, then only count the measurement if the child reaches the conclusion of the word before breaking out into “creaky voice.”

Intrasentential Pauses

- □ turn on intensity, formants, and pitch—when there is some doubt, use visual info as a guide
- □ Story 1: measure 3rd paragraph, sentences 3 and 7; 4th paragraph, sentence 1 —measuring short sentences because 1) they don’t have commas; 2) readers shouldn’t be pausing in these sentences unless they are struggling with fluency
- □ Story 2: measure 3rd paragraph, sentence 2; 4th paragraph, sentences 2 and 4 —measuring short sentences because 1) they don’t have commas; 2) readers shouldn’t be pausing in these sentences unless they are struggling with fluency
- □ look for 100 ms of both spectrographic silence and listen as well—breaths DO count as part of a pause, so silence won’t be visible, but aural cues will be there—we found that we could measure pauses of this length or greater with decent

reliability. Plus, in listening to the text, we found that pauses of this length are noticeable and would probably be rated by teachers as pauses.

- □ Sharp drops in intensity as well as nonexistent or scattered formant markers are indicators of pausing.
- □ hesitations and pre-articulation count as pausing
- □ skipped words do not count as pauses in and of themselves
- □ if a child says a phrase/word just fine, then pauses and repeats the phrase, this does not count as a pause if it's at the end of the sentence. It DOES count as a pause if it's within the sentence. In this case, there would be a pause between the end of the phrase and the beginning of the next phrase as long as it was over 100ms in length.

Intersentential Pauses—the particular sentences used here were chosen because we were already measuring something there and because it's unlikely that any merging of sounds would take place between the sentences.

- □ turn on intensity, formants, and pitch
- □ measure milliseconds of pausing between sentences
- □ Story 1: measure 3rd paragraph between sentences 3-4 and 7-8; 4th paragraph between sentences 1-2
- □ Story 2: measure 3rd paragraph between sentences 2-3; 4th paragraph between sentences 1-2 & 4-5.
- □ hesitations and pre-articulation count as pausing
- □ if a child repeats the end of a sentence in order to "fix" something prosody or pronunciation related, then start measuring at the end of the repetition.
- □ if a child repeats the first word of the sentence, even though he said it correctly initially, then only count the pause from the end of the sentence to the beginning of the initial start.
- □ If a child says the first word of the sentence incorrectly and then corrects himself, count from the end of the sentence to the correct start of the sentence.
- □ If a child skips the end of sentence punctuation, says the first word of the following sentence, and then pauses and restarts the sentence, count the pause between the false start and the correct start.

Sentence-final pitch change—particular sentences here were chosen because the final two words were both single syllable words; allowing us to measure the change from the peak of the second to last word, to the end of the final word. This prevents us from measuring change within a multi-syllable final word, where final pitch change may have already begun to take place.

- □ turn on pitch
- □ use vocalic nucleus of second to last syllable as the pitch peak—measure the average pitch of the vocalic nucleus (just as you would do in measuring for intonation contour)
- □ For final pitch, measure the pitch at the end of the final word of the sentence. Avoid measuring any "tails" which only exist because of the phonetic effects of closing off a word. These "tails" do not provide information about the child's

intonation. In cases where these “tails” exist, use judgment in finding the natural end of the word.

- □ the sentence-final pitch change is the difference when subtracting the pitch peak from the final pitch.
- □ Story 1: paragraph 3, sentences 3, 4, & 8.
- □ Story 2: paragraph 3, sentence 2; paragraph 4, sentences 1, 5
- □ if a child skips the word then don’t count that data for that child. Their average can be comprised of their other sentence-final pitch change data. Mispronunciations or incorrect words can still count, but if a child corrects or repeats him/herself, use the final correction/repetition.

Intonation Contour

- □ turn on pitch
- □ measure the average pitch of the vocalic nucleus (can include sonorant consonants when the syllable is not easily separable) of each word. in multi-syllabic words, each syllable is measured.
- □ get the SD of each sentence, then get the average SD of the three sentences
- □ use the three longest sentences of the first story, and then three independent clauses which are similar in length to the sentences from the first story. This is so that any differences will not simply be an effect of greater sentence length from the more difficult passage.
    - Story #1: paragraph 3, sentences 1, 4, and 8
    - Story #2: paragraph 3, part of sentence 3 (*Because of its small size, it became known as the Tom Thumb*) & 4 (*Cooper wanted to let people know about his new machine*); paragraph 4, sentence 5 (*Then the train picked up speed and soon it was neck and neck with the horse*)
- □ skipped words = missing data
- □ if a child says an incorrect word/pronunciation instead of the correct word/pronunciation, it can still count. Since intonation in English is more dependent on sentence and grammatical structure than any lexical information, it is likely that the student would’ve used the same intonation had s/he said the word correctly.
- □ In the case of repetitions, use the final repetition. E.g., if a student says “They may sell their…may sell their crops…” use the second “may sell their” rather than the first. The reason for this is because it is likely that the student is using a more natural intonation in the second reading than the first and they are aware of any unnaturalness in their initial reading.
- □ Each student’s final score is the average SD of the three sentences

Appendix F: Study 2a Final Draft of Comprehensive Oral Reading Fluency Scale

3rd Grade: Spring

RATE & ACCURACY		EXPRESSION		
Rating	**Rate/Accuracy**	**Rating**	**Appropriate Intonation**	**Natural Pausing**
8	137+ WCPM	4	Varies pitch and pause patterns appropriately throughout sentences to communicate meaning; makes appropriate and consistent end of sentence pitch changes.	Within-sentence pauses are short and necessary to convey meaning. Between-sentence pauses are short, but natural. Unexpected pauses are rare and brief.
6	107+ WCPM	3	Varies pitch appropriately most of the time; tends to drop pitch at the end of declarative sentences. May try to correct the prosody to match the phrasing of the text after initially getting it wrong.	Has some longer pauses within and between sentences, but they only momentarily interrupt the flow of the text. Pauses seem to be used mainly to distinguish phrases and sentences.
4	78+ WCPM	2	Intonation often may be flat or not matching the meaning/phrasing of the text (though some attempt may be made); may often not end sentences with appropriate pitch changes.	Frequent pausing within sentences; may also have some lengthy pausing between sentences
2	<78 WCPM	1	Reads with flat and unnatural intonation throughout; does not mark sentence boundaries with appropriate pitch changes.	Reading is broken and effortful with numerous pauses throughout.

Rate/Accuracy: ______ **Intonation: _______** **Pause: _______**

EXPRESSION: ______

TOTAL RATING: ______ **(*Actual WCPM: _______)***

Instructions:

- This is a Criterion-referenced assessment: students receive scores based on individual performance rather than based on comparison with other students. Pay close attention to the performance descriptions and avoid judging students' reading based on characteristics that are not included in the performance descriptions.
- Listen to students read a text which, on average, takes one-minute or more using a text (at/above) the student's grade level.
- The numbers in the WCPM column should reflect the Hasbrouk & Tindal (2006) quartiles for the appropriate grade level and time of year.
- While listening to the student read, obtain WCPM by counting errors and subtracting from the WPM. Traditional CBM errors will be marked.
    - Errors—use standard Curriculum Based Measure guidelines for counting errors:
        - **omitting** a word should count as one error
        - a **reversal** counts as one error for each word that is misplaced (e.g., switching the order of two words in a text counts as two errors)
        - **skipping** a line should count as one error, and student should be directed back to read the line once it is evident that it has been skipped
        - a **mispronunciation** counts as one error each time the word is mispronounced (correct pronunciation based on the context of the sentence). **DO NOT COUNT OFF if the student self-corrects within 3 seconds**.
        - after pausing for three seconds, the student should be directed to skip the word—this is then counted as one error
        - **DO NOT COUNT OFF** for consistent articulation and dialect interference—use your professional judgment
        - **DO NOT COUNT repetitions or hesitations as errors**
        - **DO NOT COUNT OFF for inserting words**
    - If the student reads for less than or longer than one minute, simply subtract the number of errors from the number of words read, then divide that number by the number of seconds the child read. Multiply by 60. The result is WCPM.

      Example: 75 words read – 3 errors = 72 correct words
      72 words correct / 42 seconds = 1.714
      1.714 x 60 = **103 WCPM**

    - Based on the Hasbrouck & Tindal (2006) fluency quartiles, give students a score of 2 through 8 on the rating scale.
- Also listen for intonation and pausing, and rate the child's reading on the scale of 1-4 for each category. Use the rating which most closely fits the child's reading in each category. Note, for example, that this scale can account for children who read quickly and accurately, with appropriate pausing, but with poor intonation.
- Note that while a total score is obtained, sub-scale scores are important to keep track of as they diagnose where fluency problems may lie.

Appendix G: Study 2a Interrater Agreement Analyses Separated by Text

Table G1

Percent Agreement Between Raters Across Texts and Scale Dimensions

	Level 3 Text			Level 4 Text		
	Exact	Adjacent	+/- 2	Exact	Adjacent	+/- 2
Rate & Accuracy	97	100	--	95	100	--
Appropriate Intonation	62	100	--	52	90	--
Natural Pausing	43	95	--	58	95	--
Total Expression	27	82	--	35	73	--
Total Scale Score	23	78	98	37	75	92

Table G2

Separate Text Intraclass Correlation Analyses of Rater Agreement Across Texts and Scale Dimensions

	Level 3 text		Level 4 text	
	ICC	*F**	ICC	*F**
Rate and accuracy	.98	109.50	.98	86.54
Appropriate Intonation	.74	7.47	.59	5.59
Natural Pausing	.56	4.94	.67	7.01
Total Expression	.74	10.15	.69	9.39
Comprehensive score	.93	41.62	.93	47.36

*. All tests are significant at *p* < .001, *df*(59, 59)

Appendix H: Study 2a Prosody and Reading Assessment Correlations by Rater and Text

Table H1

Correlations Between Individual Raters' Expression Ratings and Prosody Measurements

	Level 3					
Prosody Variables	Expressive Intonation		Natural Pausing		Total Expression	
	Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2
Level 3 Sentence-final Pitch	.41*	.39*	--	--	.37*	.30*
Level 3 Pitch SD	.59**	.54**	--	--	.48**	.41*
Level 3 Pause Ratio	--	--	-.55**	-.72**	-.64**	-.73**
Level 3 Intersentential Pause Length	--	--	-.27*	-.33*	-0.21	-.31*
	Level 4					
Level 4 Sentence-final Pitch	.43**	.33*	--	--	.41*	.35*
Level 4 Pitch SD	.60**	.49**	--	--	.55**	.41*
Level 4 Pause Ratio	--	--	-.77**	-.77**	-.72**	-.83**
Level 4 Intersentential Pause Length	--	--	-.26*	-.22	-.14	-.22

* *p* < .05

** *p* < .001

Table H2

Correlations Between Individual Raters' Level 3 Text Ratings and Traditional Reading Assessments

	Rater 1 Total Expression	Rater 2 Total Expression	Rater 1 Total Score	Rater 2 Total Score	QRI Fluency	QRI Comprehension	TOWRE
Rater 2 Total Expression	.82**						
Rater 1 Total Score	.89**	.89**					
Rater 2 Total Score	.79**	.95**	.96**				
QRI Fluency	.77**	.83**	.92**	.91**			
QRI Comprehension	.16	.17	.22	.22	.26*		
TOWRE	.48**	.55**	.58**	.59**	.78**	.03	
WIAT-RC	.43**	.42**	.44**	.42**	.49**	.51**	.41**

Note: L3 = Level 3

*. Correlation is significant at the .05 level (2-tailed).

**. Correlation is significant at the .01 level (2-tailed).

Table H3

Correlations Between Individual Raters' Level 4 Text Ratings and Traditional Reading Assessments

	Rater 1 Total Expression	Rater 2 Total Expression	Rater 1 Total Score	Rater 2 Total Score	QRI Fluency	QRI Comprehension	TOWRE
Rater 2 Total Expression	.81**						
Rater 1 Total Score	.89**	.91**					
Rater 2 Total Score	.79**	.96**	.96**				
QRI Fluency	.71**	.88**	.89**	.92**			
QRI Comprehension	.16	.10	.17	.13	.24**		
TOWRE	.52**	.68**	.65**	.68**	.79**	.18*	
WIAT-RC	.54**	.50**	.53**	.51**	.50**	.50**	.41**

*. Correlation is significant at the .05 level (2-tailed).

**. Correlation is significant at the .01 level (2-tailed).

Appendix I: Study 2b Final Draft of Comprehensive Oral Reading Fluency Scale

This scale is designed to measure oral reading fluency by focusing on three dimensions of fluency: automaticity (rate & accuracy measured by words correct per minute—WCPM) and prosody. Prosody, sometimes described as the music of language, has been divided into two components: intonation and pausing. **Intonation** is defined as the rise and fall of pitch when speaking, usually used to convey meaning and importance. **Pausing** can be defined as a complete absence of vocalizing as well as breaks in text from repetitions, hesitations, and pre-articulation.

3rd Grade: Spring

AUTOMATICITY (Circle rating)		**EXPRESSION**			
Rating	**WCPM**	Intonation Rating	**Appropriate Intonation** (circle rating)	Pausing Rating	**Natural Pausing** (circle rating)
8	137+ WCPM	4	Makes noticeable pitch variations throughout to communicate meaning; makes appropriate & consistent end of sentence pitch changes. One or two exceptions may exist.	4	Pauses may be used to convey meaning. Between-sentence pauses are short, but natural. Unexpected pauses occur less than once per sentence on average.
6	107+ WCPM	3	Varies pitch appropriately & makes appropriate end of sentence pitch changes most of the time. Some flatness may exist, but intonation effectively communicates meaning overall.	3	May have brief unexpected pauses once or twice per sentence, but pauses seem to be used mainly to distinguish phrases & sentences. Longer pauses are rare & only momentarily interrupt the flow of the text.
4	78+ WCPM	2	Intonation is frequently flat or not matching punctuation or the meaning/phrasing of the text. The reader shows appropriate pitch variation on a few sentences, but is flat or unnatural on many others. Overall impression is that intonation does not effectively communicate meaning.	2	Frequent pausing within sentences; may also have some lengthy pausing between sentences. May pause often between phrases or 3-4 word groupings.
2	<78 WCPM	1	Reads with flat or other unnatural intonation throughout; does not mark sentence boundaries with distinct pitch changes, except occasionally.	1	Reading is broken & effortful with numerous pauses throughout. Reads primarily in groups of 1-2 words without pausing.

Scoring	**Rating**
Automaticity WCPM Rating	
Total Expression (add Appropriate Intonation + Natural Pausing Rating)	
Comprehensive Oral Reading Fluency Score (add Automaticity WCPM Rating +Total Expression)	

Instructions:

This is a Criterion-referenced assessment: students receive scores based on individual performance rather than based on comparison with other students. Pay close attention to the performance descriptions and avoid judging students' reading based on characteristics that are not included in the performance descriptions.

STEPS:

1. Make sure the numbers in the WCPM column reflect the Hasbrouk & Tindal (2006) quartiles for the appropriate grade level and time of year.
2. Listen to students read a text which, on average, takes one-minute or more using a text (at/above) the student's grade level. While listening to the student read, obtain WCPM by counting errors and subtracting from the WPM. Traditional CBM errors will be marked.
    - Errors—use standard Curriculum Based Measure guidelines for counting errors:
        - **omitting** a word should count as one error
        - a **reversal** counts as one error for each word that is misplaced (e.g., switching the order of two words in a text counts as two errors)
        - **skipping** a line: If the child is not redirected to read the skipped line, simply count each skipped word as one error.
        - a **mispronunciation** counts as one error each time the word is mispronounced (correct pronunciation based on the context of the sentence). **DO NOT COUNT OFF if the student self-corrects**.
        - after pausing for three seconds, the student should be directed to skip the word—this is then counted as one error
        - **DO NOT COUNT OFF** for consistent articulation and dialect interference—use your professional judgment
        - **DO NOT COUNT repetitions or hesitations as errors**
        - **DO NOT COUNT OFF for inserting words**
    - If the student reads for less than or longer than one minute, simply subtract the number of errors from the number of words read, then divide that number by the number of seconds the child read. Multiply by 60. The result is WCPM.

      Example: 75 words read – 3 errors = 72 correct words
      72 words correct / 42 seconds = 1.714
      1.714 x 60 = **103 WCPM**
3. Based on the Hasbrouck & Tindal (2006) fluency quartiles, give students a score of 2 through 8 on the rating scale.
4. Also listen for intonation and pausing, and circle the most appropriate rating for the child's reading using the performance descriptions in each category. Note, for example, that this scale can account for children who read quickly and accurately, with appropriate pausing, but with poor intonation.
5. Total the ratings
    a. Write the Automaticity WCPM rating in the appropriate box at the bottom of the form
    b. Add together the Intonation and Pausing ratings to obtain the Total Expression score, and write this number in the appropriate box at the bottom of the form
    c. Add together the Automaticity and the Total Expression scores to obtain the Comprehensive Oral Reading Fluency score. Write this number in the appropriate box at the bottom of the form.

Appendix J: Study 2b Interrater Agreement Analyses Separated by Text

Table J1

Percent Agreement Between Raters Across Texts and Scale Dimensions

	Level 3 Text			Level 4 Text		
	Exact	Adjacent	+/- 2	Exact	Adjacent	+/- 2
Rate & Accuracy	98	100	--	97	100	--
Appropriate Intonation	52	98	--	62	98	--
Natural Pausing	60	97	--	60	100	--
Total Expression	38	77	--	50	85	--
Total Scale Score	37	75	98	48	82	95

Table J2

Separate Text Intraclass Correlation Analyses of Rater Agreement Across Texts and Scale Dimensions

	Level 3 text		Level 4 text	
	ICC	*F**	ICC	*F**
Rate and accuracy	.99	237.61	.98	123.10
Appropriate Intonation	.68	5.17	.77	9.48
Natural Pausing	.71	6.01	.77	7.72
Total Expression	.76	7.27	.84	12.95
Comprehensive score	.94	32.76	.95	47.56

*. All tests are significant at $p < .001$, *df*(59, 59)

Appendix K: Study 2b Prosody and Reading Assessment Correlations by Rater and Text

Table K1

Study 2b Correlations Between Individual Raters’ Expression Ratings and Prosody Measurements

	Level 3					
Prosody Variables	Expressive Intonation		Natural Pausing		Total Expression	
	Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2
Level 3 Sentence-final Pitch	.36**	.21	--	--	.26*	.12
Level 3 Pitch SD	.47**	.39**	--	--	.45**	.27*
Level 3 Pause Ratio	-.62**	-.67**	-.75**	-.69**	-.72**	-.73**
Level 3 Intersentential Pause Length	--	--	-.37**	-.56**	-.32*	-.53**
	Level 4					
Level 4 Sentence-final Pitch	.36**	.20	--	--	.31*	.27*
Level 4 Pitch SD	.60**	.26*	--	--	.52**	.42**
Level 4 Pause Ratio	-.58**	-.71**	-.67**	-.72**	-.68**	-.76**
Level 4 Intersentential Pause Length	--	--	-.40**	-.40**	-.36**	-.39**

* *p* < .05
** *p* < .001

Table K2

Correlations Between Individual Raters' Level 3 Text Ratings and Traditional Reading Assessments

	Rater 1 Total Expression	Rater 2 Total Expression	Rater 1 Total Score	Rater 2 Total Score	QRI Fluency	QRI Comprehension	TOWRE
Rater 2 Total Expression	.76**						
Rater 1 Total Score	.92**	.91**					
Rater 2 Total Score	.76**	.98**	.94**				
QRI Fluency	.76**	.87**	.90**	.90**			
QRI Comprehension	.17	.15	.17	.14	.15		
TOWRE	.71**	.76**	.80**	.79**	.84**	.01	
WIAT-RC	.45**	.43**	.45**	.42**	.47**	.47**	.35**

*. Correlation is significant at the .05 level (2-tailed).

**. Correlation is significant at the .01 level (2-tailed).

Table K3

Correlations Between Individual Raters' Level 4 Text Ratings and Traditional Reading Assessments

	Rater 1 Total Expression	Rater 2 Total Expression	Rater 1 Total Score	Rater 2 Total Score	QRI Fluency	QRI Comprehension	TOWRE
Rater 2 Total Expression	.86**						
Rater 1 Total Score	.94**	.91**					
Rater 2 Total Score	.86**	.97**	.96**				
QRI Fluency	.75**	.82**	.86**	.87**			
QRI Comprehension	.38**	.24	.31*	.27*	.29*		
TOWRE	.71**	.77**	.75**	.76**	.80**	.15	
WIAT-RC	.53**	.43**	.53**	.48**	.46**	.58**	.35**

*. Correlation is significant at the .05 level (2-tailed).

**. Correlation is significant at the .01 level (2-tailed).