

THE RELATIONSHIP OF PROSODIC READING TO READING RATE
AND OTHER CONSTRUCTS OF READING ABILITY

A Dissertation
by
MARY PROCTOR HENDRIX

Submitted to the Graduate School
at Appalachian State University
in partial fulfillment of the requirements for the degree of
DOCTOR OF EDUCATION

May 2013
Educational Leadership Doctoral Program
Reich College of Education

THE RELATIONSHIP OF PROSODIC READING TO READING RATE
AND OTHER CONSTRUCTS OF READING ABILITY

A Dissertation
by
MARY PROCTOR HENDRIX
May 2013

Approved by:

Woodrow Trathen, Ph.D.
Chairperson, Dissertation Committee

Darrell Morris, Ed.D.
Member, Dissertation Committee

Robert Schlagal, Ed.D.
Member, Dissertation Committee

Omer Ari, Ph.D.
Member, Dissertation Committee

Jim Killacky, Ed.D.
Director, Educational Leadership Doctoral Program

Edelma D. Huntley, Ph.D.
Dean, Cratis Williams Graduate School

Copyright by Mary Proctor Hendrix 2013
All Rights Reserved

Abstract

THE RELATIONSHIP OF PROSODIC READING TO READING RATE AND OTHER CONSTRUCTS OF READING ABILITY (May 2013)

Mary Proctor Hendrix
B.S., Gardner-Webb University
M.A., Ed.D., Appalachian State University

Chairperson: Woodrow Trathen

The question addressed by this study was how well the results of an adapted rating scale of fluent reading would correspond to objective measures of those same readings. To that end, a trained investigating team listened to taped recordings (114) of 38 fourth- and fifth-grade students reading fourth- and fifth-grade texts to answer questions. These students were part of a larger longitudinal study in which subjects were given a complete battery of informal assessments including measures of automatic word recognition and accuracy, rate, and comprehension of contextual reading.

After listening to a recording multiple times, the team used the adjusted rating scale to evaluate the prosodic quality of the reading. The team further analyzed the oral readings by marking observed phrasal boundaries, pauses, hesitations, and other features that marked or disrupted the flow of reading. (Skilled readers observe natural and necessary grammatical boundaries with pauses and voice fluctuation. Less skilled readers may not observe these boundaries and may also disrupt flow with unexpected pausal intrusions.) This process was an objective check on the accuracy of the rating scale

judgments. Reading rate and reading accuracy were also used as additional objective measures to further verify fluency.

Results from the prosody rating scale produced distinct groups of fluent readers, from which descriptive profiles for each group were developed. In addition, statistical cluster analysis procedures were used to form fluency groups based on objective measures of reading rate, reading accuracy, and number of pauses. Discriminant function analyses revealed that all three measures predicted fluency group membership, but reading rate and pauses were much better predictors than reading accuracy. Comparisons between groups formed by subjective prosody ratings and groups formed from the cluster analyses showed a high degree of overlap and agreement, validating the prosody ratings. Correlations revealed that reading rate and pauses correlated to prosody ratings.

Results from this study suggest that rating scales can be used accurately and productively in measuring young readers' fluency and prosody. However, the cluster analyses suggest that rating scales are most robust when used to distinguish three levels (low, middle, and high) of student performance. In addition, the data reveal that online measures of oral reading rate, number of pauses, and to a lesser degree accuracy provide objective measures of fluency and prosody. These measures are less reliant on subjective interpretation and are easy to collect, especially reading rate, which proved to be the most powerful predictor of fluency in these analyses. Reading rate, therefore, is shown to be a valid and reliable proxy measure for reading fluency.

Table of Contents

Abstract	iv
Preface: Introduction and Explanation of Study	1
Chapter One: Review of the Literature	6
A Brief History of Oral Reading in the Schools	11
Cognitive Psychology and the Reading Process	15
Early Efforts to Investigate Prosody	21
Current Efforts to Assess Prosody	29
The Present Study	34
Chapter Two: Methodology	36
The Morris et al. Longitudinal Study	36
The Present Study	40
Procedures	41
Chapter Three: Results	51
Overview of Analyses	51
Distributional Normality of Dependent Variables	52
Analysis of 4-4 Data Set	52
Analysis of 4-5 Data Set	65
Analysis of 5-5 Data Set	76
Correlations and Comparison of Correlations	85
Summary of Results	88

Chapter Four: Discussion and Implications	91
Major Findings of the Study	92
Implications for Teaching and Research	94
Limitations	97
References	99
Appendix A: Descriptive Statistics for the Print-Processing and Comprehension	
Measures (Second-Sixth Grades)	111
Appendix B: Word Recognition Measure	112
Appendix C: Passage Reading Inventory (Forms A, B, C, and D)	115
Appendix D: Passage Readability. Equivalency, Hierarchical Difficulty,	
and Stability	116
Appendix E: Screening Data	118
Vita	125

Preface: Introduction and Explanation of the Study

The author of this dissertation, Mary Proctor Hendrix, died on Friday, February 1, 2013 after a 16-month battle with cancer. She was 51 years old. Two days later, over 800 mourners attended her funeral at the First United Methodist Church in the small town of Mocksville, North Carolina. Mary had been a fixture in her community over many years. She was married to her life-long companion Chris, had raised two daughters, been active in youth ministries, taught Sunday school, directed a preschool, taught Title I reading at Cooleemee Elementary School, and later taught graduate students at Appalachian State University's Yadkin County and Davie County extension sites. Mary Hendrix was an exceptional teacher—a positive force for good in her community—and, on a chilly Sunday morning in February, the people of Davie County came to pay their respects.

There was another group of adults, perhaps 50, in the church pews that morning. These were elementary school teachers from the region who had either worked with Mary in Davie County or been students in her ASU graduate courses. These teachers knew Mary as a friend and also as a reading professional whom they respected deeply.

How Mary came to teach graduate reading courses for ASU is, itself, an interesting story. She enrolled in the Reading Education M.A. program at Appalachian State in 2000. From the start, she was an outstanding student in academic courses, exhibiting an inquiring intellect and good writing skills. However, Mary really stood out in the 4-week summer practicum where she had the opportunity to teach, under

supervision, two struggling readers. Her teaching skill was evident and her enthusiasm for teaching was real and contagious. In fact, her supervising professor commented, “As I observed the tutoring lessons, I wasn’t sure who was having more fun, Mary or her student.”

In the spring of 2004, the ASU reading faculty began to offer doctoral-level courses in reading. Mary was one of eight students in our first class. She thrived in the doctoral setting, welcoming the challenge of theory and research but always wanting to see how they applied to practice. In the spring of 2005 Mary was accepted into the Educational Leadership Doctoral Program at Appalachian State University. As a doctoral student, Mary read widely and critically in the research literature and also showed a talent for academic writing. Meanwhile, our off-campus masters program in reading was growing, and we needed instructors. Given Mary’s strength in both reading theory and practice, our faculty knew that she would do a fantastic job teaching the graduate-level courses (diagnosis and practicum) at our Yadkin and Davie Counties extension sites. And so, just 6 years after being a student, herself, in these two masters courses, Mary Hendrix begin teaching them to new ASU reading graduate students. Student evaluations of her courses were outstanding, leading one senior professor to state, “They are among the best I have ever seen.” Eventually, Mary became the hub of our graduate program in Yadkin and Davie Counties. Each spring she taught the reading diagnosis course, and each summer she organized, directed, and taught the reading practicum.

The Dissertation

Not all of Mary Hendrix’s time as a doctoral student was spent reading, learning, and teaching others. She also had to come up with a dissertation or research idea. Given

her interest in reading diagnosis, Mary was drawn to the topic of *reading fluency*. In 2000, the influential National Reading Panel Report had cited fluency as one of the five pillars of reading instruction, along with phoneme awareness, phonics, vocabulary, and comprehension. And from 2000 to 2010, there was a flurry of research studies and practitioner articles on the topic. Still, experts disagreed about the definition of reading fluency; for example, did the construct include oral reading accuracy, reading speed, prosody (intonation and phrasing), comprehension, or some combination of these (Kuhn & Stahl, 2003; Pikulski, 2006; Torgesen & Hudson, 2006)?

The ASU reading faculty (and this included Mary) had a long-standing interest in reading fluency. In our diagnostic work, we routinely separated a student's word recognition, fluency, and comprehension scores, and examined the relationships among them. Furthermore, we used reading rate as a proxy for reading fluency. That is, we believed that reading rate (words read per minute), *if recorded in a reading-for-meaning context*, was an objective, efficient, and valid measure of fluency; however, not everyone agreed with our position. In fact, by 2005 reading rate had become a controversial topic in our field. Some scholars argued that reading speed was being overemphasized in the schools; that, due to accountability pressures, teachers were encouraging students to read fast at the expense of reading fluently and with understanding. This same group argued that prosody (e.g., phrasing, intonation, and expression), as opposed to rate, was a better indicator of reading fluency.

It was into this rate vs. prosody controversy that Mary Hendrix, doctoral student, stepped. In a previous research effort, several ASU reading faculty (Morris, Trathen, and Schlagal), with Mary's assistance, had collected a large number of oral reading samples

in a rural school district. As a group, we decided to carefully examine a subset of these oral reading tapes with the goal of teasing out the relationships between oral reading accuracy, reading rate, and phrasing (one element of prosody). This would be Mary's dissertation study. The aforementioned faculty would help her listen to the tapes and score the oral reading behaviors. Mary would be responsible for reviewing the relevant literature, analyzing the data, and writing up the research report.

In the fall of 2008, our research group (Mary and the three professors and on occasion another doctoral student and friend, Amie Snow) began meeting one afternoon per week for two hours. We would listen carefully to the audiotapes of fourth and fifth graders reading fourth- and fifth-grade passages, recording the number of pauses made (appropriate and inappropriate), and rating the overall fluency of each child's reading. We moved slowly, able to listen to and discuss only four or five tapes in a 2-hour period. Nonetheless, each of us appreciated the novelty and significance of what we were doing. In a collaborative way, we were learning about reading fluency first-hand. By listening to and trying to make sense of the children's oral reading behavior, we knew that we would eventually have some new and important data to share with the reading field.

Data collection continued throughout the fall of 2008 and into the spring and summer of 2009. Mary's enthusiasm for the project kept us all motivated. She analyzed her results the following fall and spring with the help of Drs. Trathen and Ari and began to write her dissertation shortly thereafter. Dr. Trathen was Mary's dissertation advisor and chair and Drs. Morris, Schlagal, and Ari were her readers. With the usual interruptions that face a mature doctoral student (i.e., parenting, teaching, being an active community member), Mary moved forward with her writing. As do most, she found the

writing challenging and sometimes difficult; still, she loved it, a sure sign of a future scholar. Mary had nearly completed her dissertation (literature review, method, results, and a discussion outline) when, in September of 2011, she was struck (but not struck down) by a serious disease. During the next 16 months, Mary, the teacher, taught us all how “to live,” confronting her illness with steadfastness and hope. Just days before entering the hospital for the last time, she made some notes about getting started back on “that discussion section.” For Mary, it was a last bit of unfinished business. However, in truth she had all but completed her dissertation study, and a good one it is.

*Darrell Morris
Woody Trathen
Bob Schlagal
Omer Ari
May, 2013*

Chapter One: Review of the Literature

Anyone who reads or listens to readers senses the importance of fluency. Teachers, who endure halting, what seems like word-by-word oral reading by some of their students, appreciate a smooth and effortless rendering of text. Parents who anxiously compare their child's oral reading to that of a more able sibling understand the significance of fluency. The child who fears reading aloud in class certainly realizes the importance of accurate, fluid reading. Researchers (Adams, 1990; Breznitz, 2006; Chall, 1983; Perfetti, 1985; Spear-Swerling & Sternberg, 1996) clearly established the importance of fluency in a model of reading development. As a construct, fluency received national attention when the National Institute of Child Health and Human Development (NICHD) released the Report of the National Reading Panel (2000) and named fluency as a critical area of reading development and research. As part of its investigation into fourth graders' reading proficiency, National Assessment of Educational Progress (NAEP) included quantitative measures and qualitative ratings of oral reading fluency for the first time in 1992 (Pinnell, Pikulski, Wixson, Campbell, Gough, & Beatty, 1995). With renewed interest, fluency can no longer be described as the neglected reading goal (Allington, 1983), but it is still an elusive one.

The elusiveness first stems from differing importance attributed to the components of fluency. Most researchers (Adams, 1990; Carver, 1990; LaBerge & Samuels, 1974; Logan, 1988; Perfetti, 1985, 2007) focus on two readily-measured components of reading fluency, word recognition accuracy and reading rate. Others (Benjamin, Schwanenflugel, Meisinger, Groff, Kuhn, & Steiner, 2013; Dowhower, 1991;

Kuhn & Stahl, 2003; Rasinski, 2003; Rasinski, Reutzel, Chard, & Linan-Thompson, 2011; Schreiber, 1980) have concentrated their efforts on ways to investigate and evaluate the third component of reading fluency: prosody. Prosody refers to the cadence and melody of speech, which reflects an understanding of the distinctive rhythm and structure of written language (Dowhower, 1991; Schreiber, 1991). A question exists, then, regarding the components of fluency. Is prosody, often termed “reading with expression,” a critical component of fluent reading, equal in importance to accurate word recognition and reading speed? Or, is it a desirable but unnecessary and occasional characteristic of certain readers’ performances?

This lack of consensus is apparent in various definitions of reading fluency. Many researchers describe fluency as the ability to read a text quickly, accurately, and expressively (Meyer & Felton, 1999; NICHHD, 2000; Rasinski, 2003). Others focus on the accurate and automatic decoding of words (Carver, 1990; LaBerge & Samuels, 1974; Logan, 1988; Samuels, Schermer, & Reinking, 1992; Torgeson & Hudson, 2006; Wolf & Katzir-Cohen, 2001). While some researchers concentrate on the prosodic dimension of fluency, noting the natural expressiveness of reading (Allington, 1983; DeFord, 1991; Dowhower, 1991; White, 1995), others separate fluency from expressiveness (Cowie, Douglas-Cowie, & Wichmann, 2002; Young & Bowers, 1995). Attempting to articulate a simple definition of fluent reading reveals a complex network of related and competing factors (Hudson, Pullen, Lane, & Torgesen, 2009).

Responding to these differences in the reading literature, Kuhn, Schwanenflugel, and Meisinger (2010) offered a comprehensive definition of fluency:

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)

This definition separates comprehension from fluency and recognizes that fluency of online print processing enables comprehension processes (Pikulski & Chard, 2005), but fluency need not include comprehension as a central element in its conceptualization. Separating comprehension from online print processing is supported by Gough's simple view of reading at the theoretical level (Gough, Hoover, & Peterson, 1996; Gough & Tunmer, 1986; Hoover & Gough, 1990). In this model, reading comprehension is a product of two separate but interrelated and necessary components: decoding (or print processing) and language comprehension. Furthermore, recent empirical data support assessing print processing separate from assessing comprehension (Morris, Bloodgood, Perney, Frye, Kucan, Trathen, Ward, & Schlagal, 2011; Morris, Trathen, Frye, Kucan, Ward, Schlagal, & Hendrix, 2013). The definition proposed by Kuhn et al. (2010) also stresses that fluency of online print processing is composed of accurate, automatic and prosodic reading both in oral and in silent reading. The similarity between oral and silent reading processes again is supported by recent empirical data (Morris et al., 2011; Morris, Trathen, Lomax, Perney, Kucan, Frye, Bloodgood, Ward, & Schlagal, 2012). Finally, as noted by Kuhn et al. (2010), levels of text difficulty have been shown to affect online reading behaviors, including measures of rate and accuracy (Morris, et al., 2011) and likely affect measures of prosody as well (Benjamin & Schwanenflugel, 2010).

However, certain aspects of this definition require additional scrutiny. Kuhn et al. (2010) suggest that oral reading provides clear measures of fluency, including: (a) word recognition, (b) rate, (c) phrasing, and (d) intonation. This definition lists both phrasing and intonation as separate components of fluency when it may be more correct to conceptualize appropriate phrasing as a single component that is *revealed* by pausing and intonation. Furthermore, these researchers suggest that rate and accuracy cannot adequately measure fluency *without* a measure of prosody, and they assert that prosody can only be measured with a spectrograph or with prosody rating scales (Rasinski, Rikli, & Johnson, 2009; Zutell & Rasinski, 1991). This last point is tested directly in the present study.

What constitutes prosodic reading? Prosody is a complex concept that comprises many constituent features, such as (a) few pausal intrusions, or inappropriate hesitations within phrases, (b) longer phrasal units, (c) syntactically appropriate phrases, and (d) appropriate pitch changes at phrase and sentence boundaries (Dowhower, 1991). Some “equate reading with expression with reading prosody” (Kuhn et al., 2010, p. 234). Yet, this linkage is problematic because reading with expression is not defined consistently among researchers. It is one thing to conceptualize reading with expression as “using prosodic features of language, such as emphasis, pitch changes, pause placement and duration, and phrasing in accord with syntactic structure so that text is translated aloud with the tonal and rhythmic characteristics of everyday speech” (Klauda & Guthrie, 2008, p. 310). It is quite another to conceptualize it as an animated rendering of text suitable for performance (Allington, 1983; Rasinski, 2003). Not all skilled reading requires this second kind of expression.

Assessing prosody presents another set of difficulties. Accuracy and rate can be measured relatively easily and reliably by scoring errors made and amount of time it takes to read a passage orally from an informal reading inventory (Morris, 2008). Conversely, prosody is more difficult to measure reliably. The first problem is that prosody is confounded with rate and accuracy when one listens to an oral reading—the three fuse to form an impression of fluency. Traditionally, rating scales (Hudson, Lane, & Pullen, 2005; Pinnell et al., 1995; Rasinski et al., 2009; Zutell & Rasinski, 1991) have been used to assess the quality of prosodic reading. Currently, the NAEP scale (Pinnell et al., 1995) and the Multi-Dimensional Fluency Scale (MFS) (Rasinski, 2003; Rasinski et al., 2009; Zutell & Rasinski, 1991) are used widely. Yet, in the case of rating scales such as the NAEP, the rating is an overall impression of the quality of the reading and prosody is not isolated from accuracy and rate. Rasinski's MFS separates dimensions of prosodic reading into four subscales: (a) expression and volume; (b) phrasing; (c) smoothness; and (d) pace. This scale provides more discrimination than the NAEP scale, but accuracy cannot be separated from phrasing and smoothness. Furthermore, what is pacing if not rate? Finally, rating scales are often imprecise, unreliable, and in some cases, inefficient for busy teachers to use in assessing students' reading prosody (Schwanenflugel, Hamilton, Kuhn, Wisenbaker, & Stahl, 2004).

Prosody is also assessed through spectrographic measures (Dowhower, 1987, 1991; Schwanenflugel, et al., 2004) and parsing tasks (Kleiman, Winograd, & Humphrey, 1979; Young & Bowers, 1995), but these measures while more reliable are not easy, efficient, or in the case of spectrographic analysis, inexpensive. One can hardly imagine a fourth-grade teacher measuring her students' oral reading with a spectrograph.

Despite the increasing interest in prosody and its recognized contribution to reading fluency, the practicality of assessing reading prosody is still poorly understood. What is needed is research testing the idea that fluency cannot be adequately measured without also measuring prosody. Are there objective online measures of reading that can adequately measure fluency? Can measures of prosody as a construct that represents fluency be simplified? Does text difficulty affect prosody and other measures of fluency? These and other questions need careful exploration by researchers. The uncertainties surrounding the notion of prosody are not new. A brief examination of the history of oral reading in American schools shows that this is a long-standing situation.

A Brief History of Oral Reading in the Schools

Historically, prosody has been judged alternately as a goal of and a hindrance to fluent reading. In the earliest days of reading instruction in this country, expressive reading was closely tied to fluent reading. Because reading materials were rare and illiteracy was high in colonial times, oral reading was an important social activity, valued both for entertainment and information (Rasinski, 2006a; Smith, 2002). Expressive oral reading was the most highly prized educational goal. Instructions detailing proper elocution, emphasis, and emotion abounded in the professional literature of the times, as can be seen in the following explanation of how students were expected to “perform” when reading aloud:

A just delivery consists in a distinct articulation of words pronounced in proper tones, suitably varied to the sense, and the emotions of the mind; with due attention to accent, to emphasis, in its several gradations; to rests or pauses of the voice, in proper places and well-measured degrees of

time; and the whole accompanied with expressive looks, and significant gestures. (Cobb, 1835, cited in Smith, 2002, p. 37)

Teachers were urged to emphasize explicit rules concerning punctuation and the cues they provided for proper elocution. Students were made to memorize long lists of rules concerning accuracy, pronunciation, pauses, and tone of voice. Smith (2002) pointed out the inferiority of early reading books, arguing that, in the zeal to develop good elocution, the literary quality of such books suffered, thus inhibiting the motivation to read. Mann (1867) agreed, saying, “Where then, too, are the rich mines of thought contained in their readers, their first-class books, and their little libraries? These they have been accustomed to consider merely as instruments, to practise [*sic*] pronunciation, emphasis, and cadence, upon” (p. 71).

Near the beginning of the twentieth century, however, the emphasis changed, and researchers began to view prosody as a hindrance to fluent reading, rather than a critical component of it (Rasinski, 2006a; Smith, 2002). Realizing that reading is more than word-calling and performance, researchers stated that an overemphasis on prosodic reading stole attentional resources from comprehension, and they criticized classrooms in which oral reading was taught and valued exclusively. Huey (1908/1968) observed:

Reading as a school exercise has almost always been thought of as reading aloud, in spite of the obvious fact that reading in actual life is to be mainly silent reading. The consequent attention to reading as an exercise in speaking, and it has usually been a rather bad exercise in speaking at that, has been heavily at the expense of reading as the art of thought-getting and thought manipulating. (p. 359)

Huey also noted prosodic oral reading interfered with the individual's "maximum rate of effective reading" (p. 361), now viewed as the component most essential to fluent reading.

The period of 1910-1925 was characterized by advances in scientific measurement, which led to an increasing amount of educational testing (Smith, 2002). The effectiveness of instructional methods was evaluated, and recommendations were made. Results highlighted the benefits of silent reading. Some scholars advocated silent reading exclusively, but most felt that both oral and silent reading had a place in the curriculum. Judd (1921, cited in Arnold & Sableski, 2006) emphasized the importance of oral reading for the beginning reader. He believed that the young learner must see and understand the connection between the spoken word and written text. However, Judd stressed the limited benefits of oral reading. He concluded that oral reading was a more laborious practice, "a menace to intelligence" (p. 111), whose worth was exhausted as the reader entered the upper grades.

Further research indicated the superiority of silent reading with respect to speed and comprehension. Parker (1921) described the findings of an early test comparing oral and silent reading:

One of the important facts that early appeared from the use of standard tests of the reading of school children was that the rate of silent reading becomes more rapid than the rate of oral reading somewhere in the middle grades. . . .In the case of most pupils, this change comes in the third or fourth grade, depending upon the natural talent of the pupil and the

methods used in teaching reading. After this point in the grades, most pupils will read more rapidly silently than orally. (p. 262-263)

Researchers urged teachers to reorganize their instruction to reflect these new findings.

Judd (1921) stated:

As soon as we understand the character of silent reading and oral reading, we are able to organize our classroom work on a sound scientific basis. At the present time, we are in the beginnings of this organization. Enough analysis of these two kinds of reading has been made to send the word abroad among educational people that they will have to make over the school program in reading. Enough interest has been aroused to make this one of the richest and most influential fields of inquiry. (p. 662)

At this point, oral reading no longer served the cultural function it enjoyed in the nineteenth century and earlier. Because literate individuals and materials were both more plentiful, it was no longer necessary to gather in groups to hear oral reading for entertainment or information.

Silent reading offered an instructional benefit as well. All students, rather than one student reading orally, could engage in silent reading simultaneously, which resulted in more students reading a greater volume (Rasinski, 2006a). Professional books and instructional materials favored silent reading in the classroom. At the beginning of the twentieth century, then, elocutionary rules disappeared from basal readers. Silent reading became the preferred form of instruction, and prosody was largely ignored. For a time in the 1920s, teachers and students in the upper elementary grades embraced silent reading

and emphasized automaticity and comprehension. Oral reading was relegated to beginning readers in the primary grades (Smith, 2002).

However, this trend began to fade in the last half of the twentieth century. Teachers, not wholly comfortable with an approach that excluded oral reading, felt unsure about what was actually happening during silent reading. They began to implement round-robin reading, in which students took turns reading aloud an unrehearsed passage (Rasinski, 2006a). Because it is easily managed, the practice of round robin reading quickly became commonplace in the elementary classroom, although research has never supported its effectiveness. Henderson (1981) noted that “manuals of the time deplored this practice but, interestingly, teachers did it anyway!” (p. 19). In this mode of instruction, fluent reading became closely associated with accuracy. Correctness was paramount, even at the expense of fluid renderings.

Cognitive Psychology and the Reading Process

In the late 1960s in the United States, the discipline of cognitive psychology emerged and began to influence the study of reading. In stark contrast to behaviorism, with its measurable observations of surface events, cognitive psychology sought to understand the inner workings of the mind (Pearson, 2002). In effect, cognitive psychologists who studied reading in the latter part of the twentieth century, accepted the famous challenge Huey put forth in 1908:

And so to completely analyze what we do when we read would almost be the acme of a psychologist's achievements, for it would be to describe very many of the most intricate workings of the human mind, as well as to

unravel the tangled story of the most remarkable specific performance that civilization has learned in all its history. (1908/1968, p. 6)

Of particular and related importance are LaBerge and Samuels' (1974) theory of automatic information processing and Perfetti's (1985) theory of verbal efficiency during reading.

LaBerge and Samuels' theory of automatic information processing. LaBerge and Samuels (1974) developed a model of information processing in reading, which takes into account selective attention, decoding, and comprehension. Selective attention is defined as the ability to concentrate mental energy on a certain feature of the environment while withholding mental energy from other features. An individual has a limited amount of attention, or cognitive resources, available for any given task; cognitive resources that are used for one task are necessarily unavailable for another. Complex skills consist of several subskills. If each subskill requires attention, the task as a whole cannot be accomplished because the attentional requirements will be too high. However, if enough of the component skills can be processed automatically, the whole task can be completed because the demands on attention are within a tolerable range. Automatizing component skills allows one to chat while eating lunch. If chewing each bite required one's attention, there would be no cognitive resources available for meaningful conversation.

This ability to automatize certain component skills also allows one to engage in reading. Reading is a complex task, yet it can be conceptualized as two discrete but interrelated skills: decoding (or print processing) and comprehension (Gough et al., 1996; Gough & Tunmer, 1986; Hoover & Gough, 1990; Samuels, 1988). In turn, each of these skills comprise a set of interrelated subskills. Decoding refers to the ability to translate

the written text into spoken words. While the definition is simple, the ability is not. It consists of a set of complex subskills including feature detection, letter recognition, knowledge of spelling patterns, and whole word recognition (LaBerge & Samuels, 1974; see also Adams, 1990; Rayner & Pollatsek, 1989; Share, 1995). Decoding skill can be described by two levels of proficiency: accuracy and automaticity. “While accuracy of [word] recognition is a necessary and desirable goal, it is not a sufficient condition for skilled reading. . . .In order to become a fluent reader, word recognition must be both accurate and automatic” (Samuels, 1988, p. 759). Beginning readers are neither accurate nor automatic in word recognition skills. However, good instruction and extended practice will lead first to accuracy and then to automaticity. Until reading becomes automatic, beginning readers must devote most or all of their attention to decoding, leaving little cognitive energy for comprehension. With practice, however, decoding requires less mental effort, and more attention is available for comprehension.

Comprehension involves building meaning from decoded material. Readers use their background knowledge and personal experiences to interpret and understand written text. While decoding can become automatic through extended practice, comprehension always requires attention (LaBerge & Samuels, 1974). LaBerge and Samuels acknowledged that, whereas the automaticity model shows how print is sequentially processed to the point of comprehension, the model has little to say about comprehension itself. “The complexity of the comprehension operation appears to be as enormous as that of thinking in general” (LaBerge & Samuels, 1974, p. 320).

In summary, LaBerge and Samuels’s automaticity model rests on two assumptions. Although the human brain can only deal with a limited amount of

information at one time, reading with understanding requires two demanding cognitive tasks: decoding and comprehension. Therefore, the subskills involved in decoding must become automatic to allow mental resources to be available for the exacting task of comprehension.

Perfetti's verbal efficiency theory. Perfetti's (1985) verbal efficiency theory (VET) also addressed proficient reading by detailing how print processing and comprehension are related. VET offers a model of reading skill that emphasizes efficiency of linguistic processing powered by the quick execution of lower level processes. In other words, successful comprehension of text depends on the efficient operation of local processes, such as lexical access, in which words are recognized and linked to familiar concepts in the reader's memory.

Perfetti referred to verbal efficiency in terms of product and cost. "Verbal efficiency is the quality of a verbal processing outcome relative to its cost to processing resources" (p. 102). A verbal processing outcome can be any segment of the reading process, from letter identification to comprehension of a text unit. Any component skill of the reading process can be more or less efficient according to this definition. For instance, lexical access, the process of locating a written word in long-term memory, can be automatic or it can be arduous. The automation of this particular subskill offers the most potential for reaching a uniform high level of efficiency. Furthermore, if this lower level process reaches a high level of efficiency, it enables further reading processes, which depend first on lexical access, to increase in efficiency as well.

Automaticity in lexical access (Perfetti, 1985) develops through repeated encounters with words in text. An initial encounter with a particular word allows for a

partial word representation. A partial representation allows the word to be accessed, but the process requires a drain on resources, which are consequently unavailable for other processes. With practice afforded through beneficial instruction and wide reading, the number of encounters with a particular word increases, and the partial word representation is refined to a complete word representation, allowing for accurate and automatic lexical access. In this case, resources previously needed and used to identify the word are freed for other processes necessary for skilled reading. Therefore, an individual's automatic word knowledge is an important measure of reading ability.

The work of the aforementioned cognitive psychologists cemented accuracy and automaticity as necessary components of fluent reading. A third, more cautiously embraced component of fluency is the appropriate use of prosodic features. Disregarded for most of the 20th century, prosody has once again captured the attention of reading researchers, beginning with Peter Schreiber, a linguist at the University of Wisconsin.

Schreiber's prosodic cue theory. Schreiber (1980) argued that fluent reading requires more than accurate and automatic word recognition; the reader must also group words into syntactically appropriate and meaningful sets. He echoed Cromer's (1970) contention that a reader's failure to read in meaningful units indicates the tendency to process text at the word level primarily and to lose important distinctions carried across combinations of words.

Schreiber and Read (1980) noted that children especially are reliant on prosodic cues in speech. This sensitivity to prosody can be seen in an infant's vocalizations. Even before they can utter words, babies babble in intonational and temporal patterns that resemble their native language. The ability of children to learn to speak a second

language with native pronunciation further confirms the affinity they have for prosody.

Schreiber (1987) suggested that children are more dependent on prosodic cues than adults in processing the syntax of spoken sentences.

A reliance on prosody in spoken language may have implications for written text. Schreiber (1987, 1991) proposed that the reader must realize that written text lacks the prosodic cues of speech and compensate for this deficiency by using the prosodic cues that are available and supplying those cues that are not. He further contended that problems in this area may account for the difficulty some students have in moving from simple decoding to fluent reading.

In related work, Kleiman et al. (1979) explored the notion of parsing text into meaningful syntactic units. Fourth-grade students, with below-average and above-average reading skills, were asked to divide sentences at syntactic boundaries. Half of the sentences were presented in both written and spoken form, while half of the sentences were presented in written form only. Kleiman et al. found that below-average readers had difficulty parsing sentences that were presented in written form compared to sentences that were presented in written and spoken form. The students with above-average reading skills did not experience difficulty with either task. These findings supported Schreiber's contention that good and poor readers may be differentiated, in part, by their ability or inability to group written words into meaningful sets.

Schreiber reinstated prosody to relevance when thinking about reading fluency, yet with a turn away from the performance focus that dominated thinking in the nineteenth century. Instead, prosody in his theory is seen as evidence that important underlying syntactic and comprehension processes are in operation as a reader makes

meaning of a text. In fact, these fluency processes are in operation whether a reader is engaged in oral or silent reading. “It is impossible to understand a written text until we assign to it a prosody—whether we take it in silently or read it aloud” (Quirk, Greenbaum, Leech, & Svartvik, as cited in Dowhower, 1991, p. 173). From this perspective, prosody is associated with fluent reading in a new and important way.

Early Efforts to Investigate Prosody

Researchers continued to study prosody in differing ways. Clay and Imlach (1971) undertook a broad examination of prosodic features, including pauses, stress, and pitch. Others (Kowal, O’Connell, O’Brien, & Bryant, 1975) concentrated on the temporal aspects (rate and pauses) of oral reading. Some researchers separated the notions of fluency and expressiveness (Cowie et al., 2002; Young & Bowers, 1985) and examined the effect of text difficulty on these variables (Young & Bowers, 1985). Dowhower (1987) examined the effects of repeated reading on the prosodic performance of young readers. Koriat, Greenberg, and Kreiner (2002) provided insight into the nature and role of prosody in skilled reading.

In one of the earliest efforts to quantify prosodic indicators, Clay and Imlach (1971) studied variables of juncture, stress, and pitch among seven-year-old readers to determine patterns of reading prosody at different stages of reading development. The researchers described juncture as “a pause in the continuous flow of oral reading” (p. 135). Juncture ranged from brief pauses within and between words to longer pauses marking the end of sentences and major phrasal units. The authors noted that long pauses also accompanied difficulty or uncertainty in print processing.

In this study (Clay & Imlach, 1971), students were asked to read a series of

passages that were leveled in difficulty. The readings were audiotaped to allow for repeated listening and analysis. A single rater, trained in descriptive linguistics, listened to each taped reading and coded a transcript of the text according to perceptual impressions of pausing, stress, and pitch. In addition to these prosodic ratings, the students were also grouped into quartiles (high, high-middle, low-middle, and low) based on more objective measures of reading ability—accuracy and rate. Profiles, based on the prosodic behaviors of the group members, were developed for each of these quartiles.

Regarding juncture, the high reading group read with fewer pauses, shorter pauses, and longer phrases (7.4 words per phrasal unit) than the other groups. The high reading group tended to pause at punctuation but occasionally read the text without honoring these signals. The low reading group read with many more pauses than the punctuation or syntax indicated, and these pauses were much longer in duration. Not surprisingly, the phrasal units for the low reading group were much shorter (1.3 words per unit). The reading performance of the other groups (high-middle and low-middle) completed the continuum bounded by the high and low reading groups.

Kowal et al. (1975) examined how proficiency affects rate and pauses in oral reading. The researchers shaped their study around the assertion that temporal features of speech and oral reading can “serve as indicants of underlying cognitive processes” (p. 549), an idea that Schreiber’s theory embraces (1980, 1987, 1991). Reading performances of fourth- and second-grade students were compared as examples of proficient and less proficient readers, respectively. Sixty-four participants read a simple paragraph, roughly placed by the Flesch-Kincaid Grade Level readability formula at the mid-first-grade level (1.4). Using spectrographic evidence, the researchers (Kowal et al.,

1975) supported Clay and Imlach's (1971) earlier findings and determined that more proficient readers had fewer and briefer pauses than less skilled readers. Proficiency also allowed for faster reading rates and increased phrase length.

The Kowal et al. study (1975) provided additional insight into the nature of pauses. The researchers found that length of pauses, as opposed to frequency of pauses, is more variable and subject to a number of processes that may or may not be directly related to the reading task. They concluded that frequency of pauses is indicative of syntactic structure while duration of pauses may be more related to passage content.

Dowhower (1987) examined how increased proficiency in a single group of children affects reading performance. Seventeen second-grade students with good decoding skills but low reading rates participated in a repeated reading intervention lasting seven weeks. Dowhower used spectrographic evidence, a visual display of the participants' oral reading, to examine the students' performance before and after the intervention. Rate, accuracy, and prosodic reading improved significantly after the repeated reading intervention. With respect to pausing, increased proficiency led to fewer inappropriate pauses and longer phrasal units.

Adding an interesting consideration to the understanding of reading behavior, Young and Bowers (1995) investigated the effect of text difficulty on reading fluency and expressiveness in 40 average and 45 poor fifth-grade readers. Teacher ratings of each student's reading ability and performance on a standardized reading comprehension test (MacGinitie & MacGinitie, 1989) were used to designate each participant as a poor or average reader. The children's oral reading was measured on texts of increasing difficulty (second-, third-, and fifth-grade reading levels). Researchers investigated a number of

variables, including oral reading accuracy (proportion of words read correctly), rate (words per minute), prosody (expressiveness), and text phrasing. Prosody was evaluated by two trained raters using Allington's (1983) fluency scale, and knowledge of phrasal units was assessed by a parsing task, in which students marked places in the text where they would pause if reading the selection aloud.

In addition to affirming previous studies showing the oral reading of average readers to be faster, more accurate, and more fluent and expressive than poor readers, the study (Young & Bowers, 1995) also showed that average readers were more successful in parsing text. In addition, the study showed that for both groups rate and accuracy decreased as text difficulty increased.

Like Young and Bowers (1995), Cowie et al. (2002) sought to separate the notions of fluency and expressiveness. In the first part of the study, trained raters assessed the oral reading of 67 children (aged eight to ten) according to fluency and expressiveness, using two separate rating scales. The reading selection, consisting of 14 sentences and containing text features such as lists, questions, and quotations, was developed to elicit expressive reading. The authors did not specify the reading level of the passage, but the Flesch-Kincaid Grade Level readability formula placed it at a second-grade level (2.0). There was a pronounced asymmetry in the relation between fluency and expressiveness. High expressiveness was closely related to high fluency, but low expressiveness was divided fairly evenly over the three levels of fluency. Cowie et al. (2002) elaborate, "The natural interpretation is that fluency permits expressiveness, that is, it is difficult to read expressively unless one has a sufficient level of fluency, but quite possible to be fluent and to read inexpressively" (p. 53).

To investigate further the interactions between fluency and expressiveness, the researchers (Cowie et al., 2002) examined the oral reading of a balanced subsample of 24 students from the original group. Only students with high and middle fluency ratings were included in the second phase of the study; students with low fluency ratings were excluded because of the low number of participants with this attribute. High, middle, and low expressiveness scores were used to form groups. The result was four students in each group: (a) high fluency – high expressiveness, (b) high fluency – middle expressiveness, (c) high fluency – low expressiveness, (d) middle fluency – high expressiveness, (e) middle fluency – middle expressiveness, and (f) middle fluency – low expressiveness. Objective prosodic indicators, including pause structures and pitch changes, were examined with spectrographic evidence and compared to categories formed by the fluency and expressiveness ratings obtained during the first phase of the study.

Spectrographic analysis revealed that rating categories were effectively predicted by prosodic indicators. Specifically, speech rate, measured in time per syllable, differentiated fluent ratings, while pitch variations differentiated expressive ratings. Cowie et al. (2002) also noted that the study provided “objective support for conclusions that a sensitive observer might draw by listening to readers with varying levels of reading skill” (p. 76). Spectrographic analysis revealed quantifiable acoustic support of prosody ratings regarding fluency and expressiveness.

Koriat et al. (2002) examined the nature and role of reading prosody in skilled readers. Twelve students from the University of Haifa read 12 sentences containing 15-16 words under three conditions. In the unpracticed condition, they read each sentence as soon as it appeared on a computer screen; this oral reading was recorded for further

evaluation. In the practiced condition, they read the same sentences used in the previous condition four times; only the fourth reading was recorded for later evaluation. In the arbitrary condition, they were asked to read the same sentences four times, but they were told to read them according to punctuation marks, which were placed arbitrarily within the sentences. Only the fourth reading was recorded. The participants were asked to read all sentences clearly and with appropriate intonation, as if they were reading text for broadcasting. Scoring involved 12 raters that were asked to judge each recorded sentence on the basis of how natural they sounded on a scale from one (very low) to ten (very high).

The raters found that the sentences read under the unpracticed condition sounded as natural as those read under the practiced condition (Koriat et al., 2002). The prosody rating for the unpracticed condition (7.38) did not differ statistically from the prosody rating for the practiced condition (7.95). The results suggested that natural reading prosody can be produced on the first reading of an unfamiliar sentence. The results were consistent across different readers, different judges, and different sentences.

Koriat et al. (2002) also showed that pause structures are strong indicators of prosodic reading and syntactic bracketing. They proposed that reading prosody, as indicated by pause patterns, reflects a syntactic structure that provides a frame on which meaning is then applied and processed. Sentences that share syntactic structures share similar pause patterns, while sentences with different syntactic structures yield different pause patterns. The researchers offered the following two sentences for consideration:

1. The windy horse was singing when the books ate the house and saw all the justice.

2. The little girl was sleeping when the burglars entered the house and took all the jewelry. (p. 271)

A skilled reader would read both sentences with similar prosody because the sentences share syntactic structure. The semantic mismatch of Sentence 1 does not interfere with the ability to read with appropriate phrasing.

Koriat et al. (2002) hypothesized that pause patterns are produced online on the basis of structural cues before complete semantic integration is attained. If the hypothesis is correct, interference with structural information should impair readers' abilities to produce distinct pause patterns, but interference with semantic information should have a minimal effect on readers' abilities to produce clear prosodic patterns. Coherent, intact sentences were manipulated structurally and semantically to provide variants that were read aloud by participants and analyzed spectrographically. Participants were instructed to read each sentence as soon as it appeared on the computer screen; they were also told that even though some of the sentences were irregular, they should read them naturally, as if they were normal sentences.

Findings (Koriat et al., 2002) showed that interfering with sentence structure resulted in less distinct pause patterns. Removing function words (e.g., the, with, and to) and disrupting word order produced unclear pause structures. Conversely, sentences that maintained structural integrity but forfeited semantic coherence showed pause patterns that were as distinct as those produced by reading sentences that were both structurally and semantically intact.

The researchers maintained that prosody is linked to structure, is relatively indifferent to meaning, and can be produced without practice. A skilled reader would be

able to read Sentence 1 above skillfully and naturally, even without practice. Koriat et al. (2002) proposed that the ability to bracket appropriate syntactic units occurs before and may assist the processing of meaning. They held that in reading, as in speech, structure prepares for and assists meaning, stating “reading prosody discloses an intermediate representation of a sentence, one that follows structure analysis but precedes more complete semantic analysis” (p. 272).

In summary, research examining prosody and pausing in particular established several important understandings. Proficient readers have a higher reading rate than poorer readers (Clay & Imlach, 1971; Dowhower, 1987; Kowal et al., 1975; Young & Bowers, 1995), indicating that fluency is tied closely to temporal aspects of reading performance (Cowie et al., 2002). Proficient readers read with fewer pauses (Clay & Imlach, 1971; Dowhower, 1987; Kowal et al., 1975), shorter pauses (Clay & Imlach, 1971; Kowal et al., 1975), fewer inappropriate pauses (Clay & Imlach, 1971; Dowhower, 1987), longer phrasal units (Clay & Imlach, 1971; Dowhower, 1987; Kowal et al., 1975;) and greater expressiveness (Young & Bowers, 1995). Researchers have demonstrated that fluency and expressiveness are different variables (Cowie et al., 2002; Young & Bowers, 1995), and high fluency ratings are not dependent on high expressiveness ratings (Cowie et al., 2002). Natural prosody can be produced online, without practice (Koriat et al., 2002), and prosody has more to do with syntax than semantics (Koriat et al., 2002; Kowal et al., 1975). Finally, an examiner can listen to and rate oral reading, and those ratings can be supported by objective measures (Cowie et al., 2002).

Current Efforts to Assess Prosody

Prosody has become accepted as an indicator of reading processes that constitute fluent reading (Rasinski, 2006b; Rasinski et al., 2011); however, a major concern is the reliable assessment of prosody. Current efforts to assess prosody have centered on spectrographic evidence and the use of rating scales (Kuhn et al., 2010).

Spectrographic evidence. Spectrographic analysis involves converting an audiotaped oral reading into a digital visual display that can be examined according to prosodic elements, including pause structures and pitch changes. Three recent studies have attempted to situate the role of prosody in reading development, using spectrographic evidence (Miller & Schwanenflugel, 2006, 2008; Schwanenflugel et al., 2004). In addition to a common purpose, each study featured young students (grades one through three) and similar procedures. Objective measures of reading performance, such as rate and decoding ability (TOWRE; Torgesen, Wagner, & Rashotte, 1999), were used to group students into skill levels. Spectrographic analysis was used to examine the students' intonation and pausing when reading aloud. These prosodic profiles were considered for each of the groups formed by objective measures of reading ability. One criticism of these studies is that students were instructed to read the text as rapidly as they could, which may lead to an artificially fast reading that sacrifices meaning for speed. This is in contrast to instructions used in other studies (Morris et al., 2011; Morris et al., 2012; Morris et al. 2013; Samuels, 2006), in which students are directed to read as they normally do with the expectation of discussing the passage after reading it.

In general, the findings indicate that good readers have fewer pauses (Schwanenflugel et al., 2004) and poor readers have more inappropriate pauses (Miller &

Schwanenflugel, 2006). In addition, the studies found no support for a connection between comprehension and prosody, as shown by pause structures (Miller & Schwanenflugel, 2006; Schwanenflugel et al., 2004), but findings do reveal a strong relation between decoding ability and prosody (Miller & Schwanenflugel, 2006, 2008; Schwanenflugel et al., 2004). Finally, the studies showed that frequency of pauses, as opposed to pause duration, is an important indicant of prosodic reading (Miller & Schwanenflugel, 2008). Kowal et al. (1975) also noted that frequency of pauses is a purer measure of the ability to read in syntactic units, whereas pause duration is highly variable and subject to a number of processes which may or may not relate to the reading task.

Rating scales. Traditionally, rating scales have been used to evaluate prosodic reading. The most widely used rating scale, developed by the National Assessment of Educational Progress (NAEP) (Pinnell et al., 1995), sought to describe qualitatively the oral reading of fourth-grade students. The NAEP scale lists four levels of expressive oral reading. Level 4 readers preserve the author's syntax by reading in large phrases with expression. Level 3 readers read in three to four word phrases while preserving most of the author's syntax, but there is little or no expressive reading. Level 2 readers read in two-word groupings, which awkwardly alter syntax. No expressive reading is present. Finally, Level 1 readers read word by word. The reading does not preserve the author's syntax, and no expressive reading is evident (Pinnell et al., 1995).

The NAEP scale was designed to provide teachers and researchers a quick means to accurately measure the overall prosodic quality of an oral reading. Although it is used extensively, the NAEP scale has been criticized. Strecker, Roser, and Martinez (1998) pointed out that defining criteria are not identified at all levels of proficiency. For

instance, Level 1 in the NAEP scale relies solely on phrasing issues, while Level 4 references phrasing, expression, and the preservation of the author's syntax.

An alternative to the NAEP scale came from Rasinski and Zutell (Rasinski, 2003; Rasinski et al., 2009; Zutell & Rasinski, 1991). Maintaining that prosody demands a more precise means of assessment than the NAEP scale, Rasinski and Zutell developed the 16-point Multidimensional Fluency Scale (MFS). The MFS consists of four subscales designed to measure a student's expression and volume, phrasing, smoothness, and pace. Each of these areas contains four levels of proficiency. Assessed with this rubric, the most able reader will read with good volume, enthusiasm, and expression; demonstrate a good sense of phrasing, honoring clause and sentence units; provide a smooth reading with few breaks and errors that are resolved quickly; and read at a conversational rate. The least able reader will read in a quiet voice with little expression; neglect phrase boundaries and natural pitch and read in one or two word groupings; read with longer and more frequent pauses (often caused by errors); and at a slow and halting pace (Rasinski, 2003). A strength of the MFS is that it allows not only for the assessment of slow, choppy reading but also inappropriately fast reading. It should be noted, though, that pacing can be directly measured by reading rate (words per minute), smoothness can be captured by marking errors in oral reading (percentage of words read correctly), and phrasing and smoothness can be measured by marking noticeable pauses in oral reading (number of pauses). In addition, research has suggested that expressive reading is not directly related to fluent reading (Cowie et al., 2002).

Researchers continue to experiment with rating scales, attempting to develop one that effectively captures all relevant aspects of reading performance. Klaua and Guthrie

(2008) created a rating scale with five dimensions of reading fluency. The dimensions assessed included (a) pace, (b) smoothness, (c) word expressiveness, (d) phrasing, and (e) passage expressiveness. Passage expressiveness represents a new level of assessment. This dimension assesses students' "oral interpretation of the passage as a whole, including the appropriateness and consistency of the mood or tone created by their oral reading" (p. 314). A score of 1 means the oral reading performance suggested no tone or mood. A score of 2 means the student read about one-quarter of the passage in an expressive way, while a score of 3 means the student read one-half to three-quarters of the passage with a consistent, expressive tone. The highest score of 4 means that the student read the entire passage, or nearly the entire passage, in an expressive way that created a mood that matched the author's intent.

To elicit readings of this type requires specific instructions. For example, the examiner is instructed to say, "Read it as expressively as you can. It's important to make it sound interesting. You don't have to read it quickly. If you come to a word that you don't know, skip it and go to the next word" (p. 314). These directions clearly favor expressiveness over rate or accuracy. The researchers reported disappointing inter-rater reliability (.70) for the new scale. To remedy this situation, they recoded the scale, collapsing it from four points to three. The recoded scale offered a modest improved inter-rater reliability (.79). The problem with the directions, the difficulty of judging passage expressiveness, and the overemphasis on expressive reading call to question the utility of this rating scale.

Some researchers (Kuhn et al, 2010; Schwanenflugel et al., 2004) have acknowledged the utility of rating scales, even as they have argued for measuring

prosody through spectrographic procedures. They agree that rating scales offer a more practical way for the classroom teacher to assess prosody, as opposed to the expertise required to use and analyze spectrographic evidence. Yet at the same time, Kuhn, Schwanenflugel, and their colleagues question the reliability of rating scales and point out that rating scales do not provide direct measures of prosody. They contend that research is needed to link spectrographic evidence to rating scales to provide diagnostic information that is both reliable and practical. Benjamin et al. (2013) report on a study that evaluated the validity of a new scale for assessing children's oral reading fluency—Comprehensive Oral Reading Fluency Scale (CORFS)—and show correlations between CORFS and spectrographic evidence, providing support for the validity of the fluency scale.

Despite growing evidence that researchers may use rating scales to assess oral reading with some measure of validity, rating scales still present other problems. Some difficulties rest with the structure of rating scales. The descriptors used in the rating scale may be too general, leading to designations that are imprecise and holistic, or conversely too specific, each resulting in the loss of important diagnostic information (Knoch, 2009; Weigle, 2002). Research suggests that training in the construction and use of rating scales is important to maintain validity, but such training is often absent or inadequate (Beswick, Willms, & Sloat, 2005; Knoch, 2009). Another significant problem is the amount of time required to apply a rating scale designation. Rasinski (2003) acknowledges this difficulty in scoring the MFS, recommending that an examiner listen to a child's taped reading multiple times to assess each aspect of fluent reading separately.

Other limitations associated with rating scales involve the raters who must evaluate an individual's performance. Research suggests that a rater's scoring may be influenced by extraneous factors, such as gender, behavior, or socioeconomic status (Beswick et al., 2005). Raters may vary in the degree to which they adhere to the rating scale descriptors (Eckes, 2008; Knoch, 2009), their knowledge of the area they are evaluating (Bailey & Drummond, 2006; Beswick et al., 2005), and their understanding and experience with a particular rating scale (Eckes, 2008). Finally, they may vary in the severity or leniency they apply to a rating scale designation (Eckes, 2008). Thus, rating scales may reveal as much about the rater as the reader. There is no getting around the fact that rating scales introduce a measure of subjectivity into the assessment of reading behavior. What seems to be needed is a relatively simple way to operationally define and assess prosody and fluency in a tangible, measurable way.

The Present Study

This study addresses the following question: How do groups formed by subjective judgments (prosody rating scales) of reading performance correspond to groups formed by more objective measures (rate, accuracy, and pauses) of reading performance? A major goal of the present study is to establish a more efficient and objective measure of prosody and fluency by accounting for the pauses a student makes while reading orally in addition to measuring reading accuracy and rate. Accounting for pauses is an appropriate tactic to include because researchers recognize pause patterns as strong indicators of prosody and syntactic structure (Dowhower, 1987; Koriat et al., 2002; Kowal et al., 1975; Miller & Schwanenflugel, 2008). Pause structures relate to three of the prosodic features identified by Dowhower (1991): (a) pausal intrusions, (b) length of phrases, and (c)

appropriateness of phrases. Accounting for pauses in oral reading also captures phrasing and smoothness, identified by Zutell and Rasinski (1991) as two of the four dimensions of fluent reading. Furthermore, a consideration of the number and appropriateness of pauses also addresses Schreiber's (1980) notion that fluent reading depends on the ability to cluster words into meaningful groups.

Another goal of the study is to compare, with respect to membership and performance, groups of readers formed by rating scale scores to groups of readers formed by more objective measures of reading behavior.

The following research steps are completed in this study:

1. Children's (fourth and fifth graders) oral readings are evaluated by using a prosody rating scale. The rating scale scores serve to organize readers into different prosody groups, and profiles are developed for each of these groups.
2. Next, fluency groups are established, based on objective measures of children's oral reading behaviors. Objective measures of oral reading accuracy, rate, and number of pauses are used to categorize students into meaningful groups—reading fluency types. The study attempts to determine which of the measures (rate, accuracy, or pauses) best predicts fluency group membership.
3. Finally, membership and performance of the prosody groups are compared to membership and performance of the fluency groups.

Analyses are conducted to explore the relations among prosody ratings, reading accuracy, rate, and number of pauses in the oral reading of fourth and fifth graders.

Chapter Two: Methodology

The present study explores ways of assessing reading fluency with prosody ratings and examines relations among these ratings and other constructs of reading fluency. Data from a previous longitudinal study of reading fluency (Morris et al., 2011) were used. Because the participating students for the present study comprised a subset of the larger longitudinal study, it is necessary to describe the methods of the previous longitudinal study (Morris et al., 2011) as well as the methods of the present study.

The Morris et al. Longitudinal Study

Participants. The participants (274 elementary school students) in the larger study were randomly selected from the eight elementary schools in Watauga County (a rural county in western North Carolina), and they represented the socioeconomic status of the community (Morris et al., 2011). Thirty-five percent (35%) of the adults in the county attained a college degree and 85% a high school degree. Thirty-one percent (31%) of the students received free or reduced lunch, and 13% of the families were below the federal poverty level. On most measures, the county reflected state averages, including educational disbursements. Compared to state averages, the county's student population did represent less racial and ethnic diversity (93% Caucasian, 3% African American, 3% Hispanic, and 1% other). However, the population was representative of the mountain region in which the study was conducted. A sample of 38 students was selected from the original study for analysis in the present study.

The Morris et al. Study (2011) examined the reading skills of students over a period of four consecutive years. Two cohorts of students were randomly selected from the second- and third-grade populations at each elementary school. The first cohort was assessed from second through fifth grade, while the second cohort was assessed from third through sixth grade. The students were assessed in the following areas: (a) orally reading grade-leveled passages (data were oral reading rate, accuracy, and comprehension), (b) silently reading grade-leveled passages (data were silent reading rate and comprehension), (c) recognizing isolated words from grade-leveled lists (data were timed and untimed accuracy scores), and (d) spelling grade-leveled words (data were accuracy scores of spelling). Portions of the assessment were audiotaped, including oral reading. These oral reading samples (fourth- and fifth-grade levels) were the primary data source in the present study.

Assessment tasks. In the Morris et al. Study (2011) individual word list reading and passage reading assessments were administered to the students for four successive years during the months of February and March. Approximately 75 minutes split into two sessions were required to individually administer the assessments to each participant. The first session consisted of word recognition in isolation and oral reading, while the second session consisted of silent reading and spelling. The assessments were administered by a research team of university-based reading educators and carefully trained graduate students; I was a member of that research team. Means and Standard Deviations for each assessment task are reported in Appendix A.

Word recognition. The word recognition test consisted of ten 20-word lists featuring words from early first grade through eighth grade. The lists were developed by

randomly sampling the grade-level lists in *Basic Reading Vocabularies* (Harris & Jacobson, 1982). Analyses showed the lists to be reliable and hierarchically graded for difficulty across two dimensions: word frequency (Zeno, Ivens, Millard, & Duvvuri, 1995) and orthographic complexity (Morris et al., 2011). Appendix B displays the grade-level word lists analyzed in the present study and reports on the consistency, stability, and hierarchical difficulty of the WR-t (word recognition-timed) measure.

Administration. The word recognition test began with the first word on the preprimer list. The examiner opened and closed two blank cards to reveal and cover the words. The examiner opened the cards for approximately one half second to reveal the word clearly and completely. The examiner then quickly closed the cards to hide the word, completing the timed presentation of the word. If the student read the word correctly and with no hesitation, the examiner moved to the next word, repeating the procedure of revealing and quickly covering the word. The examiner moved down the list of words until the child missed a word. When the child misread a word, the examiner reopened the cards to allow the student to decode the word. This was the untimed presentation of the word. After the child's untimed response, the examiner continued timed presentation of the list words until the child required another untimed presentation. Testing stopped when the student made eleven or more errors on the timed presentation of a 20-word list.

Scoring. The student earned a timed score (WR-t) and an untimed score (WR-ut) on each word list administered. The timed score (0 - 100%) represented the percentage of words identified correctly on the timed presentation. The untimed score (0 - 100%) represented the percentage of words identified correctly on the timed presentation plus

the percentage of words identified correctly on the untimed presentation, when more time was allowed for decoding (Morris, 2008). Only the timed score (WR-t) data from the Morris et al. Study were used in the analyses of the present study.

Oral reading. There were four forms of the passage reading inventory (A, B, C, and D). Each form contained passages featuring readability levels for first grade through eighth grade. The passages were taken from trade books and from well-known, commercially available reading inventories. The reading passages were chosen for their interest value. The Morris et al. Study (2011) established the reliability, stability over time, and hierarchical difficulty of the reading passages, as well as the equivalence of the four forms. See Appendix C for a list of the passages and Appendix D for comparisons of the passages analyzed in the present study.

Administration. The materials were counter balanced. In the first year of the study, each student was randomly assigned a form of the passage reading inventory to read orally. In following years, each student read different forms of the inventory, never reading the same passage twice.

The oral reading began at the highest grade level at which the student scored 80% or higher on the word recognition-timed (WR-t) assessment. The examiner provided brief instructions and a short introduction before asking the student to begin reading. The instructions included a request that the child read the passage as he or she normally would and a notification that the examiner would ask a few questions about the passage after the child finished reading. [Note: These instructions are quite similar to a recommendation by Samuels (2006) but are fundamentally different from instructions provided by Miller and Schwanenflugel (2008) when they asked students “to read the

passage as quickly and as well as they could” (p. 341).] The examiner timed the student’s reading and marked errors on a transcript. If the child hesitated on a particular word, the examiner allowed three seconds before supplying the word. After the child finished reading, the examiner noted the time and asked five or six questions about the passage.

The child continued to read passages of increasing difficulty until he or she reached a frustration level. Signs of frustration included an oral reading accuracy score below 90%, extremely slow or disfluent reading, and an oral reading comprehension score below 50% on two successive passages.

Scoring. Three scores were obtained from a child’s oral reading. Oral reading rate (RATE) is the number of words read per minute. Oral reading accuracy (ACCURACY) is the percentage of words read correctly. Oral reading comprehension is the percentage of questions answered correctly. Notifying the students that they would be required to answer comprehension questions ensured that they read the passage in order to understand it. The mean score of 84% correct for oral reading comprehension obtained during the Morris et al. Study (2011) indicated that the children were reading for meaning and that they did comprehend the reading selections. Oral reading comprehension scores were not considered further in the present study.

The Present Study

Participants. For the present study, a smaller set of audiotapes was assembled from the larger data set collected in the Morris et al. Study (2011). Tapes obtained during the third and fourth years of the longitudinal study served as the data source and targeted fourth- and fifth-graders reading fourth- and fifth-grade level texts. Selection criteria for the present study included each fourth-grader from the original study reading one fourth-

grade level passage (4-4) and one fifth-grade passage (4-5), then, a year later, these same students reading a fifth-grade passage while in fifth grade (5-5). Participants for the present study were further restricted based on their reading the same (matched) passages. Two of the passage forms for fourth grade (4A and 4C) were randomly selected for the present study. This random selection yielded 46 audiotapes. Three tapes were excluded because these students did not read a fifth-grade passage in their fourth-grade year. Two additional tapes were excluded because the students did not read a fifth-grade passage in fifth grade. Finally, three tapes were excluded because portions of the tapes were damaged or inaudible. Ultimately, data resources included the following: (a) thirty-eight audiotapes of fourth graders reading a fourth-grade passage, (b) thirty-eight audiotapes of the same fourth graders reading a fifth-grade passage, and (c) thirty-eight audiotapes of the same children one year later (now fifth graders) reading a fifth-grade passage.

Procedures

Rating prosody in oral reading. I along with three professors of Reading Education (our research team) measured prosody by listening carefully to the 114 reading protocols (38 students, 3 passages each). Our research team listened to each audiotape multiple times. The first time, we listened without a transcript of the text in order to gain a holistic appreciation of each particular reading. Our team rated the reading for fluency, using the four-level NAEP rating scale as the basis for evaluating prosody. However, it quickly became apparent that the scale was inadequate in describing what we heard in the students' oral readings. Level 1 of the NAEP scale characterizes oral reading as word-by-word, but such reading was not observed in the present research. Even the least able readers did not read in a "word-by-word" manner, so no participants received the lowest

NAEP designation of Level 1. Furthermore, there were very few readers who received the highest designation. Level 4 of the NAEP scale describes oral reading as expressive, but we listened to very few readings that could be described in this way. Moreover, there were examples of students trying to read expressively, in which the emphasis on expression tended to interfere with the overall reading performance. For instance, a student attempting expressive reading might repeat sections in order to improve the expressive rendering.

With no readers receiving a NAEP Level 1 designation and very few receiving a Level 4 designation, most of the readers in the current study fit the NAEP descriptions of Levels 2 and 3. Yet, across these two levels, there were indeed differences in prosodic reading performances. As our team listened to the tapes, distinctions of “low 2,” “high 2,” “low 3,” and “high 3” became necessary to capture the range of readers in the data. The problem stemmed from the failure of the NAEP criteria to describe accurately the characteristics of the different reading performances that we heard on the tapes. Thus, we decided to abandon the NAEP scale and develop a rating scale that would represent more accurately the range of reading fluency that we found on the tapes.

The rating scale developed by our research team yielded five prosody levels or groupings, with Level 1 representing the least prosodic reading and Level 5 representing the most prosodic reading. Descriptions distinguishing Levels 1, 3, and 5 (low, middle, and high, respectively) were determined and served as anchor levels to guide the rating process. It was difficult to articulate descriptions for Levels 2 and 4. Because prosodic reading exists on a continuum, these intermediate levels represented differences of greater or lesser prosodic skills than the anchor levels. A Level 2 reader, for example,

was one who read more capably than a Level 1 reader but not as ably as a Level 3 reader. Used this way, the three anchor descriptions allowed for a high degree of inter-rater agreement in rating the students' prosody in oral reading, using the 5-level rating scale.

The following descriptions distinguished three anchor levels of reading PROSODY. [Note: Words written in capital letters will be used to signify any variable that will be subject to statistical analysis.] High PROSODY readers read almost as well as adults. Phrasal units were large and meaningful, and pauses occurred almost exclusively at appropriate phrase boundaries. The oral reading consistently maintained the syntax of the sentences and enhanced the listener's comprehension.

The middle PROSODY group read mostly in 3-4 word phrases. These readers signaled most phrase boundaries, providing clear syntactic units to assist the listener's comprehension. Pauses and repetitions did, at times, disrupt the flow of the reading and the syntax of the sentences. This type of oral reading seemed average for a fourth- or fifth-grade child.

Members of the low PROSODY group read in shorter phrase units, usually in 2-3 word groupings. Inappropriate hesitations and repetitions continuously disrupted the syntax of the sentences. The quality of the reading tended to hinder the listener's comprehension of the message. These oral readings clearly seemed disfluent.

Our team's descriptions of the anchor levels differed from the descriptions found in the NAEP study (Pinnell et al., 1995) but fit our data better. Each team member used these anchor descriptions to rate each oral reading protocol on a 1-5 scale; then we shared scores and checked for agreement. Any differences were discussed and settled in conference. It was often necessary to listen to a single passage multiple times to rate the

oral reading with certainty. The scoring resulted in a prosody rating of each student's oral reading of a passage (PROSODY), ranging from 1 (low) to 5 (high). Twenty Five percent (25%) of the cases were randomly selected to check for inter-rater agreement. The percentage of absolute agreement calculated for prosody ratings was .92, indicating a very high degree of agreement between the raters.

Coding pauses in oral reading. Each child's oral reading was also coded to reflect the number and placement of pauses made during reading. Before beginning this process, it was necessary for the members of the research team to align impressions of what constitutes a pause in oral reading. Several passages, not included as part of the data set, were used as "practice trials" to allow members to coordinate expectations involving the duration of a pause. While exact durational times were not established in this study (or any other, for that matter), the members of the research team were able to achieve a high level of agreement as to the nature of a pause.

After assigning a score based on the five-level rating scale, our research team listened to the oral reading again. Each member marked points in the transcript where the student paused when reading the text. Again, it was often necessary to listen to a single passage a second time to code pauses with confidence. After coding the transcript, team members compared the number and placement of pauses. Agreement was checked, and differences were settled in conference. These analyses resulted in a "number of pauses" score for each oral reading protocol. Twenty-five percent (25%) of the cases were randomly selected to check for inter-rater agreement for frequency of pauses and also was found to be very high with .95 in absolute agreement.

Standardizing pause counts. Because each oral reading selection contained varying numbers of pauses and all students did not read the same selection (e.g., some read 4A and others 4C), it was necessary to standardize the “number of pauses” scores prior to statistical analysis. The minimal number of pauses for each oral reading passage was determined by asking adult readers to note obligatory pauses in the passages. Only pauses that were unanimously selected were included in the minimal number of pauses. These minimal values were then subtracted from the total number of pauses committed by each student for the corresponding reading selection, resulting in a standard number of pauses count (PAUSES) for each reading protocol. If a child had many more pauses than the standard number of pauses for a particular passage, it indicated that he or she paused more often than was required by punctuation. If a child’s number of pauses approached the standard number for a particular selection, it indicated that he or she read more like an adult, reducing pauses to places in the text where they were signaled by punctuation. [Note: There were no PAUSE scores lower than the minimal number of pauses determined for each passage.]

Imputing missing data. Data for this study include 38 participants, each with three data sets (4-4, 4-5, and 5-5), resulting in 114 potential data records for each of the variables: (a) oral reading accuracy, (b) oral reading rate, (c) oral reading pauses, (d) oral reading prosody rating, and (e) WR-t score for a total of 570 data records. Only two data records from the total were replaced; replacement data accounted for less than 0.3% of the entire data set. In two cases (Cases 2 and 3) in the 4-5 data set, students did not complete all the sections of the reading assessment that were required in the present

study. The missing data were two scores for timed word recognition (less than 2% of WR-t scores) that needed to be imputed.

Group means were used to adjust the missing data for a student. The mean difference from one year to the next was calculated for timed word recognition, and this amount was subtracted from the student's last recorded score for the variable in question. Using the mean difference is a conservative method of imputing missing data.

Analyses. The oral reading protocols were coded and data were used to form groups (levels of reading fluency performance) based on prosody ratings as well as more objective measures of reading performance. Each data set (4-4, 4-5, or 5-5) was analyzed separately.

First, prosody ratings were used to organize students into PROSODY groups. Descriptive data were developed for each PROSODY group, showing the means and standard deviations for RATE, number of PAUSES, and ACCURACY. Trends associated with these variables were noted.

However, prosody ratings represent a rather subjective measure of reading fluency. Even when rating scales are employed carefully, consistently, and with a high degree of inter-rater agreement, the decisions rely on human judgment and some uncertainty may be unavoidable. To address this uncertainty and to verify group membership determined by prosody ratings, a second method of grouping, governed solely by observable behaviors and statistical analysis, was conducted. Objective measures (oral reading ACCURACY, reading RATE, and number of PAUSES) were collected for each reading protocol. These measures were examined through cluster analysis, a statistical technique that compares cases with regard to their pattern of scores

on the objective measures and sorts them according to the similarity of their profiles (Cross & Paris, 1988; Hammett, Van Kleeck, & Huberty, 2003; Wade, Trathen, & Schraw, 1990). The CLUSTER program of SPSS was used. In this study, agglomerative hierarchical cluster analysis was employed to classify cases, and the squared Euclidean distance between objects served as an index of dissimilarity (Everitt, Landau, Leese, & Stahl, 2011).

In agglomerative hierarchical cluster analysis, each case initially comprises a single-member cluster. At the second step, the two most similar cases are linked to form a cluster. At the next step, a third case is considered. If the third case is more similar to either the first or second case than a fourth case, it joins the first cluster. If it is more similar to the fourth case, these two cases form a new cluster. The process continues as more cases are added to existing clusters, new clusters are created, and clusters are linked to make increasingly large and dissimilar groups. In the final step, all cases are connected to form a single cluster (Kojic-Sabo & Lightbown, 1999; Norusis, 2010).

The first step in cluster analysis provides no insight because each cluster contains a single member. The final step also fails to inform because all cases, regardless of differences or similarities, are combined into a single cluster. It is the task of researchers to determine the optimal number of meaningful clusters, in which the members of a particular cluster share a strong association with each other and a weak association with members of other clusters. A large number of small clusters may result in an overly detailed interpretation of the groups, while a small number of large clusters may result in the loss of important distinctions. “There is no straightforward procedure or mathematical criterion by which unequivocal decisions could be made as to when to stop clustering”

(Kojic-Sabo & Lightbown, 1999, p. 184). Some insight is gained by checking the value of the distance statistic used to form the clusters. When this value becomes relatively large, it indicates that heterogeneous groups are being forced into a single cluster (Norusis, 2010).

For the data in this study, cluster analysis suggested an appropriate number of FLUENCY groups and showed which cases, or reading protocols, belonged to each group. Members of each FLUENCY group should be associated more strongly with each other than with members of other groups. The next step in the analysis is to determine the profile for each cluster. Cluster analysis reveals structures in the data on mathematical grounds, but it does not explain these structures. Researchers must characterize each cluster based on the data examined. In this study, an online contextual reading episode was captured in terms of reading RATE, number of PAUSES, and ACCURACY. Readers were clustered into groups on the basis of their performance according to these variables, and the mean scores of the three variables were calculated and compared for each of the clusters.

The resulting FLUENCY groups for each data set were tested (using MANOVA procedures) to determine if the groups were statistically different from each other. Then, if found to be statistically different, discriminant analysis was used to determine which variable or combination of variables best predicted FLUENCY (cluster group) membership.

Analysis of Variance (ANOVA) is appropriate for situations in which more than two groups are simultaneously compared on dependent variables (Stevens, 2002). This technique was used to examine differences between groups (based on prosody rating

scores) on dependent measures of reading behavior (RATE, PAUSES, and WR-t). To determine where the significant differences lie, the Tukey post-hoc procedure was conducted. The Tukey procedure provides clear and meaningful comparisons while restricting both Type I and Type II errors (Stevens, 2002).

Lastly, correlations between the prosody ratings and oral reading variables (ACCURACY, RATE, and number of PAUSES) were examined. Schatschneider & Lonigan (2010) state that “Correlations are a measure of the association between two variables, and they can be computed on any set of paired variables regardless of the distributions or variance properties of those variables. . . . [s]tatistical tests for correlations (and means) are robust against violations of normality” p. 348. Comparisons of these correlations (e.g., the relation of prosody ratings and PAUSES versus the relation of prosody ratings and RATE) provided additional information about the relative strength among all variables.

Reading performance measures. Performance measures of reading serve as repeated dependent measures in the analyses. The measures are:

1. ACCURACY—derived from calculating the number of word-level errors made while reading and subtracting that from the total number of words read. The resulting number of words read correctly was then divided by total number of words read, resulting in a percentage correct rendering of a reading episode. This dependent variable was used in most analyses.
2. RATE—derived from calculating words per minute of a reading episode. This dependent variable was used in all analyses.

3. PAUSES—derived from counting the number of pauses in a reading episode and calculating a standard pause count for an oral reading episode. This dependent variable was used in all analyses.
4. WR-t—derived from calculating the percentage of words read correctly from a graded-list of words, where the words are presented in a timed condition. This dependent variable represents automatic word recognition ability that is independent of contextual reading variables. This variable was used in the final ANOVA analyses.

Group measures. Group measures serve as independent, grouping variables in the analyses. They include:

1. PROSODY group membership – derived from listening to the oral reading protocols and scoring prosodic fluency based on the five-level fluency scale. This independent variable was ultimately reduced to three levels (groups) and ANOVA procedures were used to test for differences between the PROSODY-RECODED groups.
2. FLUENCY group membership – derived from cluster analyses where objective measures (reading accuracy, reading rate, and number of pauses) of the same reading protocols used to score PROSODY were examined to form FLUENCY groups. These groups were tested for differences and discriminate analyses were used to determine the variables that best predict FLUENCY group membership.

Chapter Three: Results

Several issues surround the concept of prosody and its role in reading fluency assessment. What is the most efficient way to measure prosody? Must an assessment of reading fluency include a measure of prosody? Or, are there objective, measurable reading behaviors that can capture fully the assessment of reading fluency without including a prosody measure? The present study begins an examination of these issues by addressing the following questions: How do groups formed by subjective judgments (prosody rating scales) of reading performance correspond to groups formed by more objective measures (rate, accuracy, and pauses) of reading performance? What are the relations among these variables?

Overview of Analysis

This chapter reports on the results of the statistical analyses conducted on the three data sets: (a) 4-4, (b) 4-5, and (c) 5-5. The same procedures were applied to each of the data sets. PROSODY groups were formed by prosody rating scale scores; FLUENCY groups were formed from statistical analyses of RATE, PAUSES, and ACCURACY.

The prosody rating scale yielded five levels of readers (PROSODY), with accompanying qualitative descriptions (see Chapter Three). To verify the results of the PROSODY ratings obtained by the research team, cluster analyses were used to group readers on the basis of objective measures of reading ability (RATE, PAUSES, and ACCURACY). These cluster analyses favored a solution forming three FLUENCY groups instead of the five groups formed by the prosody ratings. MANOVA procedures

determined that the three FLUENCY groups derived from cluster analyses were statistically different. Discriminant function analyses were used to determine which variable or set of variables best predicted membership in the three FLUENCY groups. Based on the strength of the three-group solution derived from objective measures, the five PROSODY groups were reduced to three PROSODY-RECODED groups to match the three-group solution proposed by cluster analysis. The ANOVA procedure was used to see how these PROSODY-RECODED groups differed on the basis of measures of contextual reading ability (RATE and PAUSES) and automatic word knowledge (WR-t). Finally, a set of correlational analyses was used to reveal the relations among PROSODY-RECODED membership, RATE, PAUSES, and ACCURACY.

Distributional Normality of Dependent Variables

Prior to analyses on the dependent variables, data sets were screened for non-normality, and results show the data sets to conform to assumptions of normality required for the statistical procedures used in this study. A complete description of the methods used to investigate non-normal distribution is provided in Appendix E.

Analysis of 4-4 Data Set

Reading PROSODY groups. Fourth-grade students' oral readings of fourth-grade passages were rated for PROSODY. As described in the Method section, our research team developed a five-level prosody rating scale that placed the 38 students in one of five PROSODY groups (Group 1 is the lowest and Group 5 the highest).

Descriptive data of reading behaviors (RATE, PAUSES, ACCURACY) for the five PROSODY groups (4-4 data) are shown in Table 1. Some clear trends are evident from the data, and consistently show lower PROSODY groups' scores falling below higher

PROSODY groups' scores on RATE, PAUSES, and ACCURACY. The mean scores for reading RATE increase steadily from Group 1 (85 wpm) through Group 5 (188.83 wpm). A reading RATE of 85 wpm, the mean score for Group 1, is considerably below the expected range (110-150 wpm) for a fourth-grade reader (Morris, 2008). The mean score for Group 3 is 135.2 wpm, which falls within the expected range, and mean scores for Groups 4 and 5 are above the expected range of reading RATES.

Table 1

Descriptive Data by PROSODY Groups (4-4 Data Set)

Variable	PROSODY Groups				
	Group 1	Group 2	Group 3	Group 4	Group 5
	$n = 2$ \bar{x}	$n = 10$ \bar{x}	$n = 10$ \bar{x}	$n = 10$ \bar{x}	$n = 6$ \bar{x}
Reading RATE	85	99.9	135.2	161.3	188.83
(SD)	(21.213)	(10.545)	(17.812)	(12.64)	(14.729)
Number PAUSES	50	45.6	32.3	25.8	23
(SD)	(2.828)	(5.42)	(4.243)	(2.043)	(4.381)
Oral reading ACCURACY	92.5	94.2	96.8	96.6	97.83
(SD)	(.707)	(2.74)	(2.25)	(2.118)	(1.94)

The scores for the number of PAUSES also follow a regular pattern. Members of lower PROSODY groups pause more often when reading than members of higher PROSODY groups. The mean number of PAUSES for PROSODY Group 1 is 50, for

Group 3 is 32.3, and for Group 5 is 23. It is interesting to note that the mean number of PAUSES for Group 4 (25.8) is very close to the mean score for Group 5.

With regard to oral reading ACCURACY, members of Group 1 had the lowest mean score (92.5), Group 2 had the next lowest mean score (94.2), while members of Group 5 had the highest mean score (97.83). However, the ACCURACY scores of Groups 3, 4, and 5 are quite similar.

Reading FLUENCY groups: Cluster analysis. To organize students into meaningful FLUENCY groups and to check the validity of the prosody ratings, the scores for reading rate (RATE), number of pauses (PAUSES), and oral reading accuracy (ACCURACY) for the 4-4 data set were examined through cluster analysis. The goal of cluster analysis is to compare cases with regard to their pattern of scores and to sort them according to the similarity of their profiles. Members of each group should be associated more strongly with each other than with members of other groups.

In the cluster analysis for the 4-4 data set, a substantial increase in the distance statistic occurred between the solutions with four and three clusters. This indicated that four clusters would provide a favorable solution. This decision was supported by evidence from the dendrogram depicting the clustering process (see Figure 1). The dendrogram begins with the 38 single-member clusters and ends with one large cluster containing all cases. The four-cluster solution (marked in Figure 1 by a cutoff line) appears valid because it occurs before the distances at which clusters are combined become too large, reflecting greater dissimilarity within the groups.

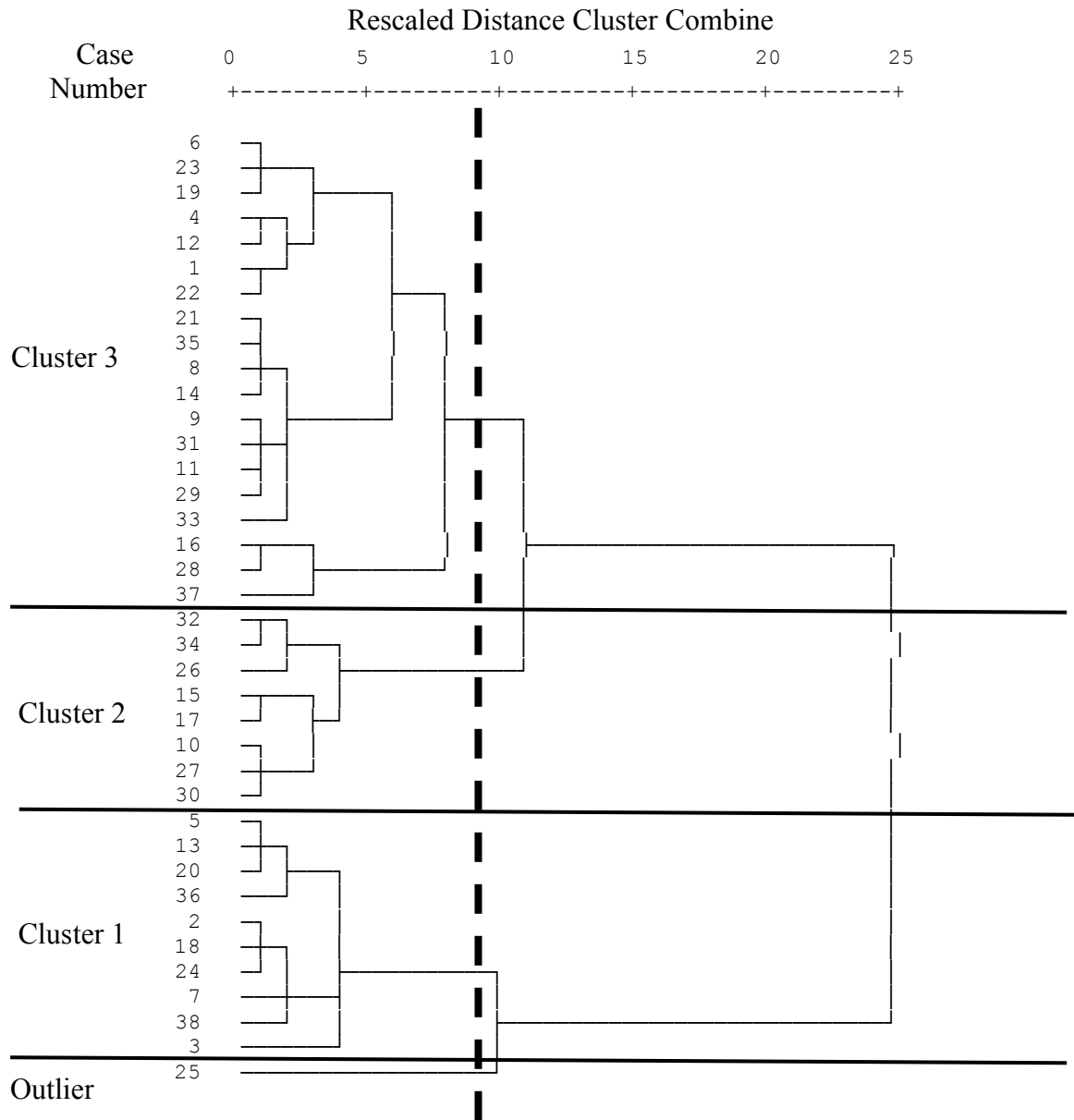


Figure 1. Dendrogram Using Average Linkage (Between Groups) for 4-4 Data Set.

However, the optimal number of clusters for this data set is three because Cluster 4 consists of only one member (Case 25). A single-member cluster provides little clarification of the relation between the examined variables of contextual reading and reading fluency. Its member is essentially an outlier whose pattern of scores did not fit easily into any of the other three clusters. Upon close examination of all the scores, it was

decided to place this Case 25 in Cluster 2 because the number of pauses and reading rate align most closely with scores obtained by other members of Cluster 2, although the accuracy score for Case 25 is the lowest in this cluster.

Cluster analysis of the 4-4 data resulted in the formation of three distinct reading FLUENCY groups: (a) Cluster 1: low ($n = 10$), (b) Cluster 2: middle ($n = 9$), and (c) Cluster 3: high ($n = 19$). Members of the low reading FLUENCY group had low reading rates, ranging from 70 – 107 wpm. Members of this group made numerous pauses (39 – 55) while reading. The accuracy scores for members of this group ranged from 92 to 95, and all readers in this group received PROSODY rating scale scores of 1 or 2.

Most readers in the middle reading FLUENCY group had higher reading rates (102-140 wpm) than readers in the low group. The number of pauses among readers in the middle group ranged from 30 to 42, and accuracy scores for members of this group ranged from 92 to 99. Most readers in this group received PROSODY rating scale scores of 3, with two readers receiving ratings of 2.

Members of the high reading FLUENCY group had high reading rates (143-218 wpm) and low numbers of pauses (17-30). The accuracy scores for members of this group ranged from 93 to 100. Most of the readers in this group received PROSODY ratings of 4 or 5; only three members received a PROSODY rating of 3.

The FLUENCY groups formed through cluster analysis and the PROSODY rating scale scores assigned by the research team share a good deal of agreement. In general, the low PROSODY ratings are assigned to members of the low FLUENCY group, and high PROSODY ratings are assigned to members of the high FLUENCY group. Most of the

readers in the middle FLUENCY group have PROSODY ratings of 3. Table 2 shows the frequency of PROSODY scores according to FLUENCY groups.

Table 2

PROSODY Rating Scores by FLUENCY Groups (4-4 Data Set)

FLUENCY Group	PROSODY Ratings
1 – Low ($n = 10$)	2, 2, 2, 2, 2, 2, 1, 1, 2, 2
2 – Middle ($n = 9$)	2, 2, 3, 3, 3, 3, 3, 3, 3
3 – High ($n = 19$)	4, 5, 5, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 3, 5, 5, 5, 5

Discriminant analysis. Following the determination of the FLUENCY clusters, discriminant function analysis was used to identify the variables that best predict group membership. First, the vectors of means of the three variables listed in Table 3 for the three reading FLUENCY clusters were compared for equality using a one-way MANOVA procedure. The resultant F -ratio of 26.595 ($df = 6, 66$), based on Wilk's Lambda (.086), indicated that the three sets of mean scores differed across the three reading FLUENCY groups significantly at $p < .001$. Group means and standard deviations for the three variables (RATE, PAUSES, and ACCURACY), as well as the associated univariate F -ratios, appear in Table 3.

Table 3

Descriptive Data and Univariate Comparisons (4-4 Data Set)

Variable	Reading FLUENCY Group			<i>F</i>	<i>p</i>
	Low (<i>n</i> = 10) \bar{x}	Middle (<i>n</i> = 9) \bar{x}	High (<i>n</i> = 19) \bar{x}		
Reading RATE	95.4 ^A	122 ^B	169.21 ^C	75.597	.000
(SD)	(12.98)	(13.702)	(18.253)		
Number PAUSES	47.5 ^A	35.33 ^B	25.36 ^B	94.859	.000
(SD)	(4.927)	(4.664)	(3.386)		
Oral reading ACCURACY	92.9 ^A	97.44 ^B	96.94 ^B	15.565	.000
(SD)	(1.101)	(2.242)	(2.068)		

Note. Within each row, means having the same letter in their superscripts are not significantly different from each other at $p < .05$.

Following the identification of a significant multivariate *F*-ratio, a stepwise discriminant function analysis was performed using reading RATE, PAUSES, and ACCURACY as predictors of membership in the three reading FLUENCY groups. For the sample, evaluation of assumptions of linearity, normality, multicollinearity or singularity were satisfactory. Homogeneity of variance-covariance matrices (Box's $M = 17.847$, Approximate $F = 1.269$, $p = .230$) was observed.

Two discriminant functions were calculated. The first discriminant function was statistically significant, $\Lambda = .086$, $\chi^2(6, N = 38) = 83.57$, $p < .001$. The second

discriminant function was also statistically significant, $\Lambda = .696$, $\chi^2(2, N = 38) = 12.317$, $p < .005$. With the use of a jackknifed (one case at a time deleted) classification procedure for the total sample of 38 readers, 100% of the cases were correctly classified. The stability of the classification procedure was checked by a cross-validation run, which was successful at correctly classifying 97.4% of the cases, with only one case misclassified. This indicates a high degree of consistency in the classification scheme.

The two discriminant functions account for 94.2% and 5.8%, respectively, of the between-group variability. As shown in the plot of group centroids in Figure 2 below, the first discriminant function maximally separates high, middle, and low reading FLUENCY groups.

Standardized discriminant function coefficients in Table 4 reveal that on the first function, the three FLUENCY groups were distinguished primarily by RATE and PAUSES. Reading rate is the predictor with the highest absolute value. Reading rate and pauses are also the two predictors that correlate significantly with the discriminant scores produced by the first function, according to the structure loadings on this function in Table 4. All loadings are in excess of .50 except for ACCURACY, whose loading on the first function is .314.

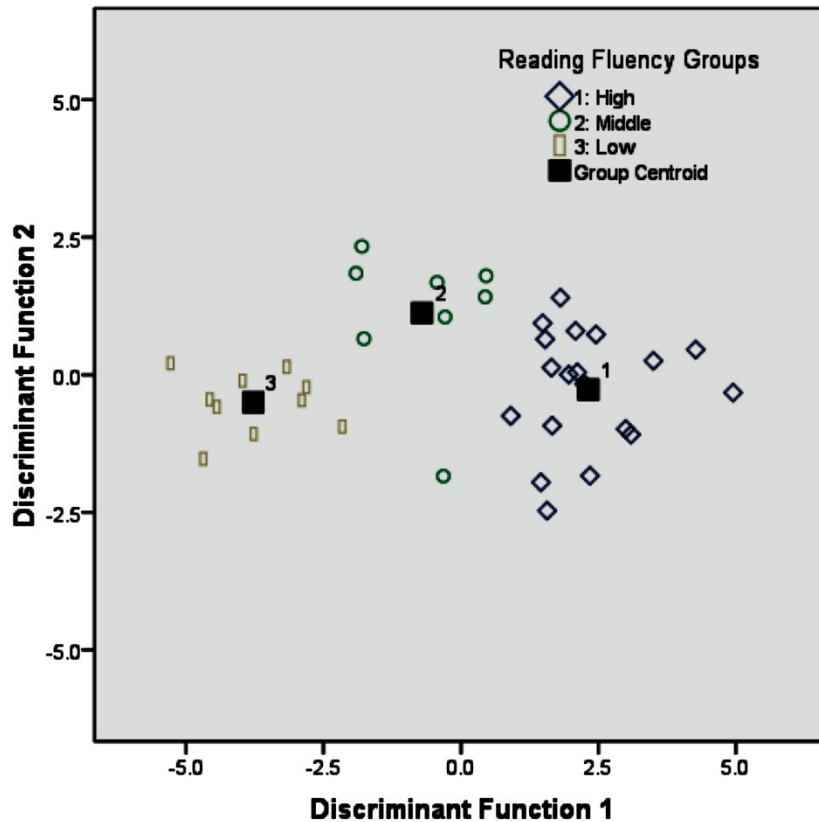


Figure 2. Plot of Group Centroids (4-4 Data Set).

Thus, the first and most important discriminant function distinguished three FLUENCY groups on the basis of high measures of RATE and low measures of PAUSES. The eigenvalue associated with the first discriminant function is about 16 times the size of the eigenvalue associated with the second function. This further emphasizes the distinction to be made between the three groups. The substantial canonical correlation of .936 between the first discriminant function and the FLUENCY group variable reflects a considerable degree of relationship between the composite of variables in the derived function and the fluency level (FLUENCY group membership) of the subjects.

Table 4

Discriminant Functions (4-4 Data Set)

Variable	Function 1		Function 2	
	Standardized coefficient	Structure loading	Standardized coefficient	Structure loading
RATE	.520	.771*	-.537	-.437
PAUSES	-.638	-.871*	-.169	-.163
ACCURACY	.141	.314	.850	.868*
Eigenvalue	7.131		.437	
Relative percentage	94.2		5.8	
Canonical correlation	.936		.551	

Chi-square (Functions 1 and 2) = 83.570 ($df = 6, p < .001$)

Chi-square (Function 2) = 12.317 ($df = 2, p < .005$)

The second discriminant function distinguishes the middle group from the other two FLUENCY groups. However, it fails to distinguish the high from the low FLUENCY group. ACCURACY is the variable that contributes most to this discriminant function, and it is the only variable that loads significantly on the function. The canonical correlation between the second discriminant function and the grouping variable (.551) reflects a considerable association between ACCURACY and the FLUENCY group of the subjects.

In summary, using cluster analysis, objective measures of contextual reading (oral reading RATE, number of PAUSES, and ACCURACY) were examined to form FLUENCY groups. A three-group solution (high, middle, and low FLUENCY) was deemed best after inspecting the cluster results. Discriminant analysis was used to determine which variable or combination of variables (RATE, PAUSES, and ACCURACY) accounted for FLUENCY group membership. Results on 4-4 data reveal that the first function of RATE and PAUSES discriminates FLUENCY group membership much better than the second function of ACCURACY: (a) RATE and PAUSES account for substantially more variance (94.2%) compared to ACCURACY (5.8%), (b) the correlation of these functions to group membership is much higher for RATE and PAUSES (.936) than for ACCURACY (.551), (c) the eigenvalue of RATE and PAUSES (7.131) is much higher than that of ACCURACY (.437), and (d) ACCURACY fails to discriminate group membership between high and low FLUENCY groups, while RATE and PAUSES clearly separate all three groups.

These data reveal that fluency can be assessed with objective measures that can reveal differences between groups. Number of PAUSES was shown to be a good measure of fluency, as was reading RATE. Furthermore, inspection of FLUENCY cluster profiles suggests that PROSODY ratings have the potential to discriminate groups. The low FLUENCY group had lowest PROSODY ratings (1 and 2), the middle FLUENCY group had mostly middle PROSODY ratings (3), and the high FLUENCY group had higher PROSODY ratings (4 and 5).

The next analysis tested the relations of PROSODY ratings and measures of contextual reading used in the classification of FLUENCY groups (RATE and PAUSES)

and a measure of automatic word knowledge (WR-t). The validity and reliability demonstrated by the three-group solution to the cluster analysis led to a recoding of the original PROSODY ratings to form three groups (PROSODY-RECODED). The characteristics of the original three anchor levels accurately describe the three PROSODY-RECODED groups. Ratings of 5 and 4 formed the high PROSODY-RECODED group, ratings of 3 formed the middle PROSODY-RECODED group, and ratings of 2 and 1 formed the low PROSODY-RECODED group. These three PROSODY-RECODED rating groups were then examined for differences.

PROSODY-RECODED groups compared: ANOVA. A one-way ANOVA was performed on the 4-4 cases to test mean differences among all three rating groups (high, middle, and low PROSODY-RECODED) on each measure of reading RATE, number of PAUSES, and timed word recognition (WR-t). Reading RATE and number of PAUSES were included because discriminant analysis demonstrated their importance and relation to FLUENCY group membership. ACCURACY was not included because it was shown to be less effective in discriminating group membership and because it was not normally distributed, violating assumptions of ANOVA (see Appendix D). Timed word recognition (WR-t) was included because it represents a valid and reliable measure of automatic word knowledge that is independent of contextual reading variables (Morris et al., 2011; Morris et al., 2012). Group means and standard deviations for the three variables appear in Table 5.

Table 5

Group Means and Standard Deviations on Dependent Variables (4-4 Data Set)

Variable	Reading PROSODY-RECODED Groups		
	Low	Middle	High
	$n = 12$ \bar{x}	$n = 10$ \bar{x}	$n = 16$ \bar{x}
RATE	97.42 ^A	135.2 ^B	171.63 ^C
(SD)	(12.866)	(17.813)	(18.913)
PAUSES	46.33 ^A	32.3 ^B	24.75 ^C
(SD)	(5.262)	(4.243)	(3.296)
WR-t	69.17 ^A	87 ^B	96.56 ^B
(SD)	(16.765)	(8.882)	(3.966)

Note. Within each row, means having the same letter in their superscripts are not significantly different from each other at the .05 level.

different means on RATE ($F[2,35] = 66.079, p < .001, \eta^2 = .791$), PAUSES ($F[2,35] = 89.356, p < .001, \eta^2 = .836$), and WR-t ($F[2,35] = 22.458, p < .001, \eta^2 = .562$).

The Tukey procedure revealed that all PROSODY-RECODED groups differed on RATE and PAUSES with all pairwise comparisons significant at $p < .001$. On WR-t, pairwise group comparisons were significant ($p < .005$) for all groups except for the middle and high PROSODY-RECODED groups ($p = .084$). The actual difference in means for WR-t between the middle and high PROSODY-RECODED groups appears to

be considerable (87 and 96.56, respectively). For the practitioner, such a difference would be important. The small number of cases included in the analysis probably resulted in the pairwise group comparison barely missing statistical significance for WR-t between the middle and high PROSODY-RECODED groups.

Analysis of the 4-5 Data Set

Reading PROSODY groups. Data from fourth-grade students orally reading fifth-grade passages were scored for prosody ratings. The PROSODY ratings used by our research team on the 4-5 oral reading protocols produced only four groups of readers with differing levels of PROSODY ratings. No readers in the 4-5 data received the highest possible PROSODY rating of 5. This likely indicates the difficulty level of the text as fourth-grade students were reading fifth-grade level text. The majority of students again are found in the middle groups; Group 2 has 15 members, and Group 3 has 11 members. Fewer students are in Group 1 ($n = 5$) and Group 4 ($n = 7$).

Certain trends are evident in the data. As in the 4-4 data set, reading RATE and number of PAUSES again provide the clearest patterns. Group 1 has the lowest mean scores for reading RATE (79.2 wpm), while Group 4 has the highest mean scores for reading RATE (153 wpm). While RATE increases with higher PROSODY groups, the number of PAUSES decreases. Members of Group 1 paused more often (57.8) than members of other groups; members of Group 4 paused less often (30.57) than members of other groups. Like reading RATE, oral reading ACCURACY increases from group to group, but the increases are not nearly so striking. The mean scores for ACCURACY only vary from 93.6 to 96. Descriptive data for these measures of reading ability are shown in Table 6.

Table 6

Descriptive Data by PROSODY Groups (4-5 Data Set)

Variable	PROSODY Groups			
	Group 1	Group 2	Group 3	Group 4
	$n = 5$ \bar{x}	$n = 15$ \bar{x}	$n = 11$ \bar{x}	$n = 7$ \bar{x}
Reading RATE	79.2	99.73	132.36	153
(SD)	(10.963)	(17.834)	(12.387)	(11.284)
Number PAUSES	57.8	45.6	36.09	30.57
(SD)	(8.642)	(5.901)	(4.826)	(8.642)
Oral reading ACCURACY	93.6	94.93	95.45	96
(SD)	(2.701)	(1.907)	(2.018)	(2.16)

Reading FLUENCY groups: Cluster analysis. The cluster analysis for the 4-5 data set presented some interesting challenges. Three cases (Cases 33, 18, and 38) resisted classification until the latest stages of cluster analysis. The group membership of these cases will be discussed later in this section.

Disregarding these three cases, the cluster analysis revealed a four-cluster prospective solution that appeared valid. However, two clusters (both exhibiting lower scores) were very similar when the descriptive data for these cases were examined; one of these two clusters contained a small number of cases ($n = 7$). The seven cases all received

PROSODY ratings of 1 or 2, and the number of PAUSES and reading RATE fit within the ranges for the other low scoring group. The only difference was these seven cases demonstrated slightly higher ACCURACY scores.

The good fit of three variables was balanced against the ill fit of one variable, and the decision was made to combine these two lower scoring groups to form the low FLUENCY group. This decision was supported by evidence from the dendrogram shown in Figure 3. The three-cluster solution (marked in Figure 3 by a cutoff line) appears valid because it occurs before the distances at which clusters are combined become too large, reflecting greater dissimilarity within the groups. Thus, it was decided to focus on three clusters, which also allowed for direct comparison to the 4-4 data set of the same students.

The three outliers were then considered. It appeared that Case 33 eventually would have been grouped into cluster 3 (high FLUENCY group). After a careful examination of the scores obtained by Case 33, it was determined that high FLUENCY group would be the most appropriate placement. The scores associated with this case include the lowest number of PAUSES for all cases, the highest ACCURACY score, and the second highest RATE. This case obviously belongs in the group of high FLUENCY readers. Cases 18 and 38 presented a different situation. These cases share very low

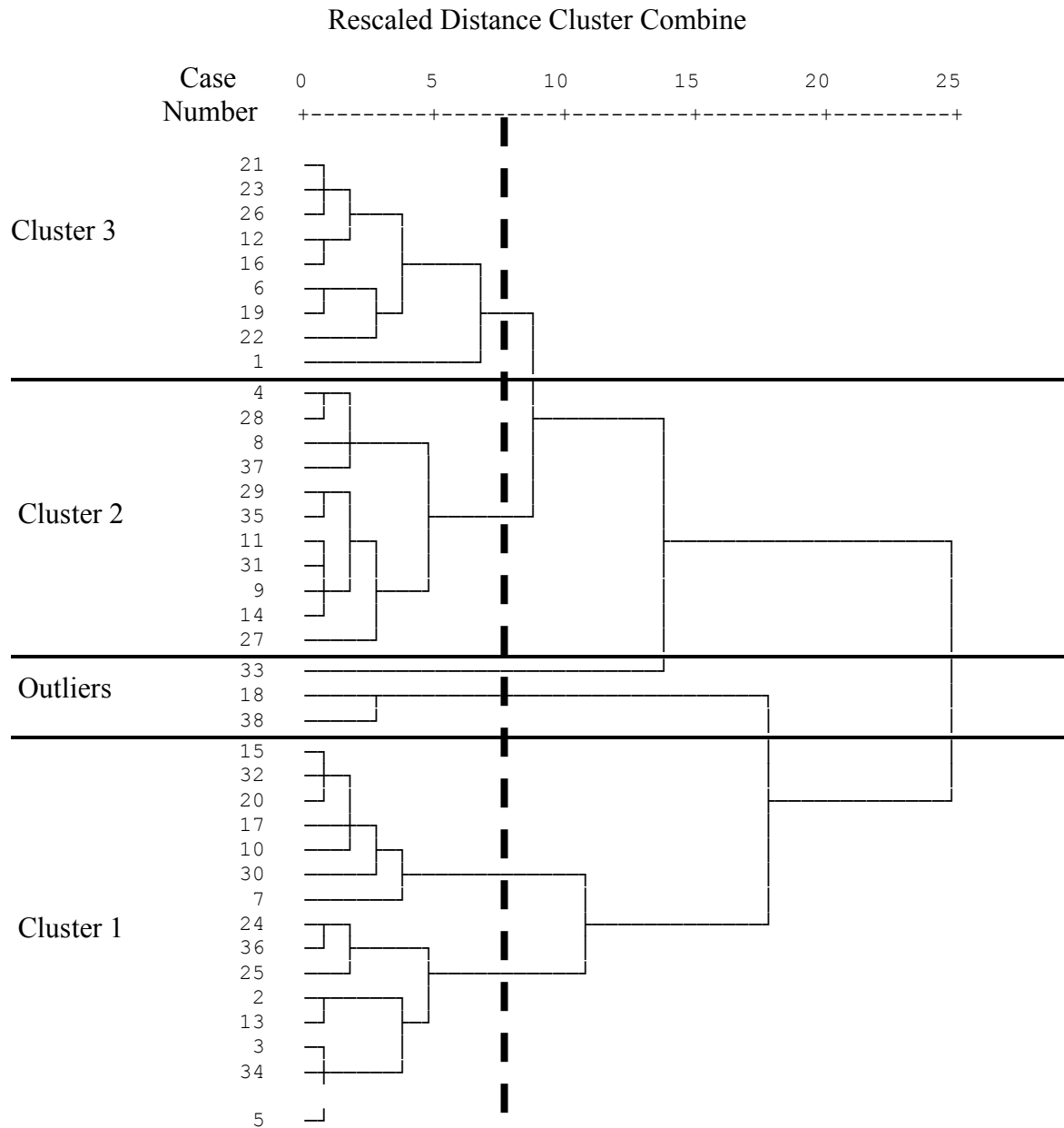


Figure 3. Dendrogram Using Average Linkage (Between Groups) for 4-5 Data Set.

RATES, a high number of PAUSES, and low ACCURACY scores. These cases belong in the low FLUENCY group (Cluster 1). The three-cluster explanation succeeded in

creating categories in which the members have more in common with each other than with members of other categories.

Cluster analysis of the 4-5 data resulted in the formation of three distinctive reading FLUENCY groups: (a) low ($n = 17$), (b) middle ($n = 11$), and (c) high ($n = 10$). The increase in text difficulty likely resulted in a shift in numbers from the high group to the middle and low groups, compared to how these students were grouped for the 4-4 data set. Members in the low group again had low reading RATES, ranging from 67 to 116 wpm, and numerous PAUSES, ranging from 40 to 70. Readers in this group received PROSODY ratings of 1 or 2, with one reader receiving a PROSODY rating of 3. The ACCURACY scores for members of this group ranged from 91 to 97.

Most members of the middle group had higher RATES (111-153 wpm) than readers in the low group. The number of PAUSES among middle group readers ranged from 34 to 43, a range quite similar to that of the 4-4 middle group (30-42). Most readers in the middle group received PROSODY rating scale scores of 2 or 3 with only two readers receiving a score of 4. The ACCURACY scores for members of the middle reading FLUENCY group ranged from 96 to 98.

Members of the high group tended to have high reading RATES (129-168 wpm) and low numbers of PAUSES (26-36). All readers in this group received PROSODY ratings of 3 or 4. The fact that no readers in the 4-5 data received a PROSODY rating of 5 is likely due to the increase in text difficulty. The ACCURACY scores for members of this group ranged from 92 to 99.

The PROSODY rating scale scores assigned by the research team and the FLUENCY groups formed through cluster analysis show general agreement. Almost all

members of the low FLUENCY group have PROSODY ratings of 1 or 2; one member received a score of 3. Members of the high FLUENCY group received ratings of 3 and 4. For the most part, the readers in the middle FLUENCY group have ratings of 2 or 3. Table 7 shows the frequency of PROSODY scores according to FLUENCY groups.

Table 7

PROSODY Rating Scores by FLUENCY Groups (4-5 Data Set)

FLUENCY Group	PROSODY Ratings
1 – Low ($n = 17$)	2, 2, 1, 2, 2, 2, 1, 2, 2, 3, 2, 1, 2, 2, 2, 1, 1
2 – Middle ($n = 11$)	2, 3, 4, 4, 3, 3, 2, 2, 3, 2, 3
3 – High ($n = 10$)	3, 3, 3, 3, 4, 4, 4, 4, 3, 4

Discriminant analysis. Following the determination of the FLUENCY clusters, the data were analyzed to identify the variables that predict FLUENCY group membership in a discriminant function analysis procedure. First, the vectors of means of the three variables listed in Table 8 for the three reading FLUENCY groups were compared for equality using a one-way MANOVA procedure. The resultant F -ratio of 18.050 ($df = 6, 66$), based on Wilk's Lambda (.143), indicated that the three sets of mean scores differed across the three FLUENCY groups at $p < .001$. Group means and standard deviations for the three variables (RATE, PAUSES, and ACCURACY), as well as the associated univariate F -ratios, appear in Table 8.

Table 8

Descriptive Data and Univariate Comparisons (4-5 Data Set)

Variable	Reading FLUENCY Group			<i>F</i>	<i>p</i>
	Low	Middle	High		
	(<i>n</i> = 17) \bar{x}	(<i>n</i> = 11) \bar{x}	(<i>n</i> = 10) \bar{x}		
RATE	89.58 ^A	128.91 ^B	147.8 ^C	62.199	.000
(SD)	(13.564)	(15.397)	(12.389)		
PAUSES	50.52 ^A	38.09 ^B	30.6 ^B	41.620	.000
(SD)	(7.706)	(2.981)	(3.272)		
ACCURACY	93.94 ^A	97 ^B	95 ^B	10.445	.000
(SD)	(2.015)	(.894)	(1.886)		

Note. Within each row, means having the same letter in their superscripts are not significantly different from each other at $p < .05$.

A stepwise discriminant function analysis was performed using RATE, PAUSES, and ACCURACY as predictors of membership in three FLUENCY groups. For the sample, evaluation of assumptions of linearity, normality, multicollinearity or singularity were satisfactory and homogeneity of variance-covariance matrices (Box's $M = 11.871$, Approximate $F = 1.802$, $p = .094$) was observed.

Two discriminant functions were calculated. The first discriminant function was statistically significant, $\Lambda = .157$, $\chi^2(4, N = 38) = 63.787$, $p < .001$, as was the second, $\Lambda = .738$, $\chi^2(1, N = 38) = 10.487$, $p < .005$. With the use of a jackknifed (one case at a time deleted) classification procedure for the total sample of 38 readers, 92.1% of the cases

were correctly classified, with misclassification of one case. The stability of the classification procedure was checked by a cross-validation run, which was successful at correctly classifying 89.5% of the cases, with two cases misclassified. This indicates a high degree of consistency in the classification scheme.

The two discriminant functions account for 91.2% and 8.8%, respectively, of the between-group variability. As shown in the plot of group centroids in Figure 4 below, the first discriminant function maximally separates high, middle, and low reading FLUENCY groups.

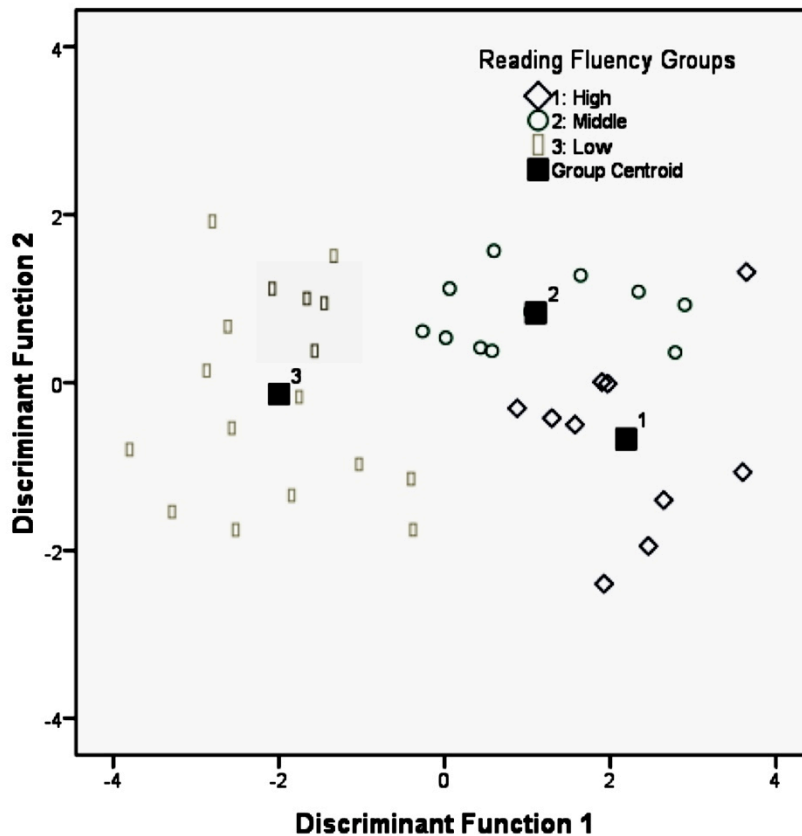


Figure 4. Plot of Group Centroids (4-5 Data Set).

Standardized discriminant function coefficients in Table 9 reveal that on this function the three FLUENCY groups were distinguished primarily by RATE. Reading RATE is the predictor with the highest absolute value, and it is the predictor that correlates significantly with the discriminant scores produced by the first function. RATE is the only variable that loads significantly on this function.

Table 9

Discriminant Functions (4-5 Data Set)

Variable	Function 1		Function 2	
	Standardized coefficient	Structure loading	Standardized coefficient	Structure loading
RATE	.966	.980*	-.270	-.200
ACCURACY	.201	.269	.982	.963*
Eigenvalue	3.688		.355	
Relative percentage	91.2		8.8	
Canonical correlation	.887		.512	

Chi-square (Functions 1 and 2) = 63.787 ($df = 4, p < .001$)

Chi-square (Function 2) = 10.487 ($df = 1, p < .005$)

Thus, the first and most important discriminant function distinguished three FLUENCY groups on the basis of high measures of RATE. The eigenvalue associated

with the first discriminant function is about ten times the size of the eigenvalue associated with the second function; this further underscores the distinction to be made between the three groups. The substantial canonical correlation of .887 between the first discriminant function and the FLUENCY group variable reflects a strong relationship between RATE and the reading FLUENCY level of the subjects.

Inspection of the plot of centroids (Figure 4) and Tables 8 and 9 suggests that the second discriminant function distinguishes the middle group from high and low FLUENCY groups. The variable that contributes most to the discriminant scores produced by this function is ACCURACY, which is the only variable that loads significantly on this function. Problematically, this function does not differentiate between the low and the high reading FLUENCY groups.

In summary, using cluster analysis, objective measures of contextual reading (RATE, PAUSES, and ACCURACY) were examined to form FLUENCY groups. A three-group solution (high, middle, and low FLUENCY) was judged best after inspecting the cluster results. Discriminant analysis was employed to determine which variable or combination of variables accounted for FLUENCY group membership. Results on 4-5 data reveal that the first function of RATE distinguishes FLUENCY group membership much more favorably than the second function of ACCURACY: (a) RATE accounts for substantially more variance (91.2%) compared to ACCURACY (8.8%), (b) the correlation of these functions to group membership is much higher for RATE (.887) than for ACCURACY (.512), (c) the eigenvalue of RATE (3.688) is much higher than that of ACCURACY (.355), and (d) ACCURACY fails to discriminate group membership between high and low FLUENCY groups, while RATE clearly separates all three groups.

With the exception of the PAUSES contributing little to the FLUENCY classification, the 4-5 data set results are comparable to the results from the 4-4 data set. Reading RATE is shown to be the best predictor of FLUENCY group membership.

The validity and reliability demonstrated by the three-group solution to the cluster analysis led to a recoding of the original prosody ratings to form three groups, just as in the 4-4 data set. Ratings of 4 formed the high PROSODY-RECODED group, ratings of 3 formed the middle PROSODY-RECODED group, and ratings of 2 and 1 formed the low PROSODY-RECODED group. These three PROSODY-RECODED rating groups were then examined for differences. The descriptive characteristics for these three groups are the same as the characteristics described in the 4-4 data set.

PROSODY-RECODED groups compared: ANOVA. A one-way ANOVA was performed on the 4-5 cases to test mean differences among all three PROSODY-RECODED groups (high, middle, and low) on each measure of contextual reading (RATE and PAUSES) and a measure of automatic word knowledge (WR-t). Reading RATE was included because discriminant analysis revealed its importance and relation to FLUENCY; number of PAUSES was included to compare 4-5 results to 4-4 results. ACCURACY was not included because it was shown to be less effective in distinguishing group membership and because it did not exhibit normal distribution, which would have violated assumptions of ANOVA. Timed word recognition (WR-t) was included because Morris, et al. (2012) found that this variable represents a valid and reliable measure of isolated, automatic word knowledge, independent of the variables associated with contextual reading. Group means and standard deviations for the three variables appear in Table 10.

Table 10

Group Means and Standard Deviations on Dependent Variables (4-5 Data Set)

Variable	Reading PROSODY-RECODED Groups		
	Low	Middle	High
	$n = 20$	$n = 11$	$n = 7$
	\bar{x}	\bar{x}	\bar{x}
RATE	94.6 ^A	132.36 ^B	153 ^C
(SD)	(18.517)	(12.388)	(11.284)
PAUSES	48.65 ^A	36.09 ^B	30.57 ^B
(SD)	(8.412)	(4.826)	(3.644)
WR-t	61.25 ^A	91 ^B	90 ^B
(SD)	(21.391)	(9.954)	(7.638)

Note. Within each row, means having the same letter in their superscripts are not significantly different from each other at the .05 level.

The one-way ANOVA resulted in significant overall difference on all measures. The analysis revealed that there were at least two groups with significantly different means on RATE ($F[2,35] = 43.057, p < .001, \eta^2 = .711$), PAUSES ($F[2,35] = 23.011, p < .001, \eta^2 = .568$), and WR-t ($F[2,35] = 14.195, p < .001, \eta^2 = .448$).

The Tukey post-hoc procedure was conducted on each measure to determine where significant differences exist and revealed that all PROSODY-RECODED groups differed in RATE with pairwise comparisons significant at $p < .05$. For the other

measures, (PAUSES and WR-t) the low PROSODY-RECODED group differed from both the middle and high PROSODY-RECODED groups with pairwise comparisons significant at $p < .05$. However, the middle and high PROSODY-RECODED groups did not differ significantly on either of these measures (PAUSES, $p = .235$; WR-t, $p = .993$).

Analysis of 5-5 Data Set

Reading PROSODY groups. Data from fifth-grade students orally reading fifth-grade passages were scored for PROSODY ratings. Prosody ratings for the 5-5 data set generated five PROSODY groups of readers. Again, the largest groups are those in the middle; combined, groups 2, 3, and 4 contain 20 members, while Groups 1 and 5 each contain only four members. Obvious patterns are evident. Reading RATE increases with group membership, while number of PAUSES decreases with group membership. Oral reading ACCURACY also increases from group to group, but because of the limited range inherent in this variable, these increases are small. The mean ACCURACY score for Group 1 is 91.25, and the mean ACCURACY score for Group 5 is 98.25. Descriptive data for these measures of reading ability are shown in Table 11.

FLUENCY groups: Cluster analysis. The cluster analysis for the 5-5 data set failed to classify one case (Case 24), so this outlier was removed from consideration. For the remaining 37 cases, a four-cluster possibility was evident from the cluster analysis. However, members of two clusters were difficult to distinguish when their characteristics were examined. Their PROSODY ratings, reading RATE, number of PAUSES, and oral

Table 11

Descriptive Data by PROSODY Groups (5-5 Data Set)

Variable	PROSODY Groups				
	Group 1	Group 2	Group 3	Group 4	Group 5
	$n = 4$	$n = 10$	$n = 13$	$n = 7$	$n = 4$
	\bar{x}	\bar{x}	\bar{x}	\bar{x}	\bar{x}
Reading RATE	86.5	100.3	125.31	166.57	185.25
(SD)	(10.503)	(16.33)	(11.101)	(18.21)	(9.287)
Number of PAUSES	63	56.5	38.84	31.57	27.25
(SD)	(10.230)	(8.181)	(7.765)	(5.126)	(5.058)
Oral reading ACCURACY	91.25	93.9	95.76	97.28	98.25
(SD)	(2.629)	(3.573)	(2.241)	(1.380)	(1.707)

reading ACCURACY scores were found to be similar. Therefore, these cases were combined to form the lower scoring FLUENCY group. Membership assignments were not violated because the cluster analysis showed the cases as eventual members of the same group. As with the other data sets, the three-cluster solution was deemed best and is supported by evidence from the dendrogram in Figure 5.

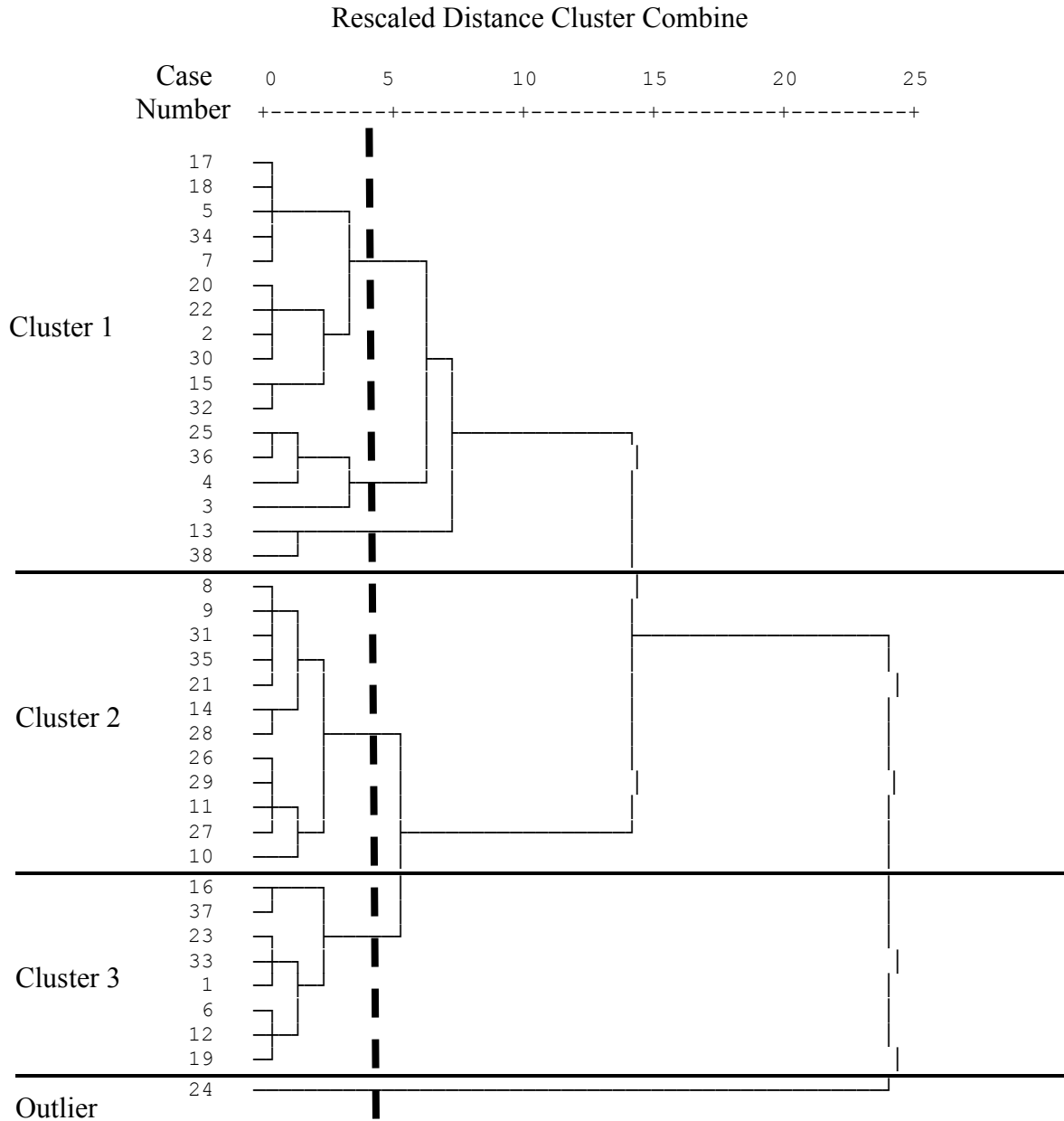


Figure 5. Dendrogram using Average Linkage (Between Groups) using 5-5 Data Set.

Cluster 1 indicates the low FLUENCY group ($n = 17$), Cluster 2 refers to the middle FLUENCY group ($n = 12$), and Cluster 3 indicates the high FLUENCY group ($n = 8$). Members in the low FLUENCY group had low reading RATES, ranging from 72 – 137 and high numbers of PAUSES (38 – 73). The ACCURACY scores for members

of this group ranged from 89 to 98. Readers in this group received PROSODY rating scale scores of 1, 2, or 3.

Members of the middle FLUENCY group had reading RATES that ranged from 115 to 152 wpm. The number of PAUSES among middle group readers ranged from 29 to 44, and ACCURACY scores ranged from 95 to 99. Readers in this group received PROSODY ratings of 3 or 4.

Members of the high FLUENCY group had the highest reading RATES, ranging from 171 – 197 wpm. They also had the lowest number of PAUSES (21 – 35). The ACCURACY scores for members of this group ranged from 95-100. Readers in this group received high PROSODY rating scale scores of 4 or 5. Table 12 shows the frequency of PROSODY rating scores according to FLUENCY groups.

Table 12

PROSODY Rating Scores by FLUENCY Groups (5-5 Data Set)

FLUENCY Group	PROSODY Ratings
1 – Low ($n = 17$)	1, 2, 2, 2, 2, 2, 3, 3, 2, 2, 2, 3, 2, 3, 1, 1, 1
2 – Middle ($n = 12$)	4, 3, 3, 3, 3, 4, 4, 3, 3, 3, 3, 3
3 – High ($n = 8$)	5, 5, 4, 5, 4, 5, 4, 4

Discriminant Analysis. The vectors of means of the three variables listed in Table 13 for the three FLUENCY groups were compared for equality using a one-way MANOVA procedure. The resultant F -ratio of 20.573 ($df = 8, 62$), based on Wilk’s

Lambda (0.075), indicated that the three sets of mean scores differed across the three FLUENCY groups at $p < .001$. Group means and standard deviations for the three variables, as well as the associated univariate F -ratios, appear in Table 13.

Table 13

Descriptive Data and Univariate Comparisons (5-5 Data Set)

Variable	Reading FLUENCY Group			F	p
	Low	Middle	High		
	($n = 17$)	($n = 12$)	($n = 8$)		
	\bar{x}	\bar{x}	\bar{x}		
RATE	101.29 ^A	132.67 ^B	182.88 ^C		
(SD)	(18.499)	(12.901)	(8.078)	79.778	.000
PAUSES	55.88 ^A	35.33 ^B	28.37 ^B		
(SD)	(10.03)	(5.757)	(4.657)	41.854	.000
ACCURACY	93.64 ^A	97.25 ^B	97.37 ^B		
(SD)	(2.498)	(1.138)	(1.685)	15.565	.000

Note. Within each row, means having the same letter in their superscripts are not significantly different from each other at $p < .05$.

A stepwise discriminant function analysis was performed using RATE, PAUSES, and ACCURACY as predictors of membership in the three FLUENCY groups.

Assumptions of normality and homogeneity (Box's $M = 23.565$, Approximate $F = 1.671$, $p = .067$) were observed.

As in the other analyses, two discriminant functions were statistically significant: The first function accounts for 79.8% of the variance ($\Lambda = .077$, $\chi^2(6, N = 37) = 84.812$, $p < .001$) and the second 20.2% of the variance ($\Lambda = .448$, $\chi^2(2, N = 37) = 26.475$, $p < .001$). A high degree of consistency in the classification scheme was observed (97.3% of the cases were correctly classified with jackknifed procedure) as well as stability (only two cases misclassified in the cross-validation). As shown in the plot of group centroids in Figure 6 below, the first discriminant function clearly separates high, middle, and low reading FLUENCY groups, while the second fails to separate the high from the low.

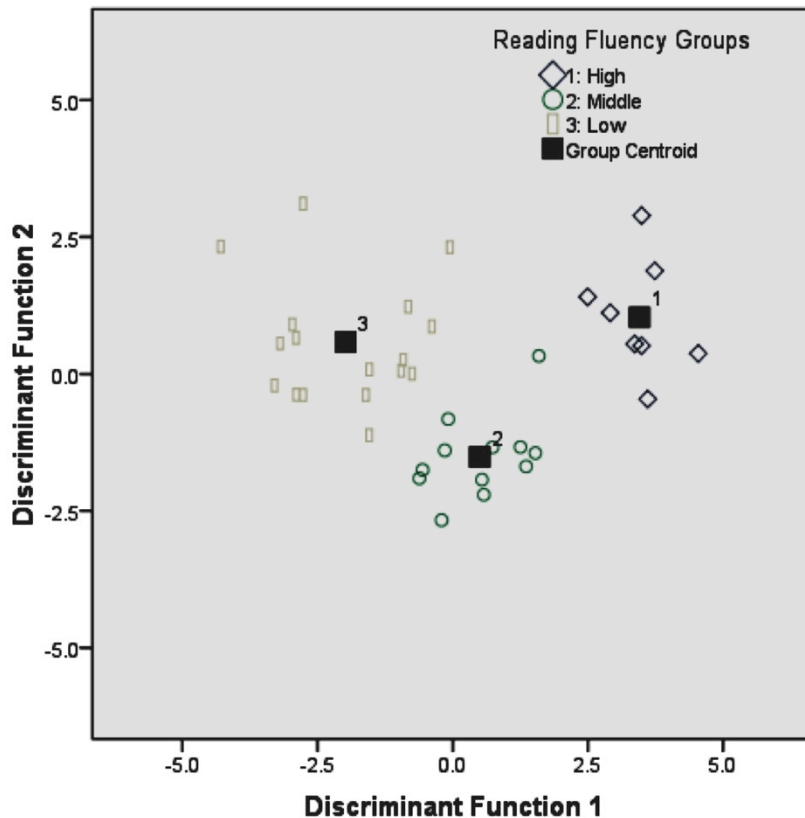


Figure 6. Plot of Group Centroids (5-5 Data Set).

Results from the 5-5 data (see Table 14) reveal that the first function of RATE and PAUSES discriminates FLUENCY group membership much better than the second function of ACCURACY: (a) RATE and PAUSES account for substantially more variance (79.8%) compared to ACCURACY (20.2%), (b) the correlation of these functions to group membership is much higher for RATE and PAUSES (.911) than for ACCURACY (.743), (c) the eigenvalue of RATE and PAUSES (4.858) is much higher than for ACCURACY (1.231), and (d) ACCURACY fails to discriminate high and low FLUENCY groups, while RATE and PAUSES clearly separate all three groups.

Table 14

Discriminant Functions (5-5 Data Set)

Variable	Function 1		Function 2	
	Standardized coefficient	Structure loading	Standardized coefficient	Structure loading
RATE	.853	.979*	1.097	.167
PAUSES	-.126	-.677*	1.193	.440
ACCURACY	.209	.380	-.701	-.416*
Eigenvalue	4.858		1.231	
Relative percentage	79.8		20.2	
Canonical correlation	.911		.743	

Chi-square (Functions 1 and 2) = 84.812 ($df = 6, p < .001$)

Chi-square (Function 2) = 26.475 ($df = 2, p > .001$)

As in the 4-4 and 4-5 data sets, these data reveal that fluency can be measured with objective measures that can reveal differences between groups. Furthermore, number of PAUSES was shown to be a good measure of FLUENCY, as was reading RATE. Again, the validity and reliability demonstrated by the three-group solution to the cluster analysis led to a recoding of the original PROSODY ratings of the 5-5 reading protocols to form three PROSODY-RECODED groups. As in the other analyses, ratings of 5 and 4 formed the high PROSODY-RECODED group, ratings of 3 formed the middle PROSODY-RECODED group, and ratings of 2 and 1 formed the low PROSODY-RECODED group, and the descriptive characteristics of each group matched the 4-4 and 4-5 data sets. These three PROSODY-RECODED groups were then examined for differences.

PROSODY-RECODED groups compared: ANOVA. A one-way ANOVA was performed on the 5-5 cases to test mean differences among all three PROSODY-RECODED groups (high, middle, and low) on each measure of RATE, PAUSES, and WR-t. Group means and standard deviations for the three variables are displayed in Table 15. At least two PROSODY-RECODED groups had significantly different means on RATE ($F[2,34] = 80.009, p < .001, \eta^2 = .825$), PAUSES ($F[2,34] = 46.056, p < .001, \eta^2 = .730$), and WR-t ($F[2,34] = 19.556, p < .001, \eta^2 = .535$).

Tukey post-hoc tests revealed that RATE differed between all groups, with pairwise comparisons significant at $p < .001$. Similarly, PAUSES differed for all groups, $p < .05$. For timed WR-t, the low PROSODY-RECODED group was significantly different from both the high and middle PROSODY-RECODED groups ($p < .05$), but the high and middle groups were not significantly different from each other ($p = .105$).

Table 15

Group Means and Standard Deviations on Dependent Variables (5-5 Data Set)

Variable	Reading PROSODY-RECODED Groups		
	Low	Middle	High
	$n = 13$ \bar{x}	$n = 13$ \bar{x}	$n = 11$ \bar{x}
RATE	95.46 ^A	125.31 ^B	173.36 ^C
(SD)	(16.148)	(11.101)	(17.71)
PAUSES	58.92 ^A	38.84 ^B	30 ^C
(SD)	(9.031)	(7.765)	(5.31)
WR-t	72.31 ^A	88.08 ^B	96.36 ^B
(SD)	(13.481)	(8.549)	(3.233)

Note. Within each row, means having the same letter in their superscripts are not significantly different from each other at the .05 level.

Correlations and Comparison of Correlations

The correlations shown in Table 16 indicate certain trends featuring RATE and PAUSES. RATE is significantly correlated with PROSODY-RECODED membership, PAUSES, and ACCURACY. A faster RATE signals a higher prosody rating, fewer PAUSES, and a higher ACCURACY score. Furthermore, the number of PAUSES is significantly correlated with PROSODY-RECODED and ACCURACY. Fewer PAUSES

Table 16

Correlations between the Prosody-Recorded Variable and Oral Reading Variables for Data Sets 4-4, 4-5, and 5-5

	4-4 Data Set ($n = 38$)			4-5 Data Set ($n = 38$)			5-5 Data Set ($n = 37$)		
	2	3	4	2	3	4	2	3	4
(1) Prosody-Recorded	.889**	-.903**	.496**	.833**	-.738**	.266	.899**	-.833**	.590**
(2) Oral Reading Rate		-.880**	.487**		-.876**	.362*		-.853**	.585**
(3) Number of Pauses			-.622**			-.324*			-.554**
(4) Oral Reading Accuracy									

Note. * $p < 0.05$. ** $p < 0.01$.

are associated with a stronger prosody rating and a higher ACCURACY score. Of the three oral reading variables, ACCURACY is correlated lowest with the PROSODY-RECODED variable in all data sets. In addition, with the more difficult text (the 4-5 data set), the correlations were not as strong as they were for the other two data sets.

Statistical comparisons of the correlations, displayed in Table 17, provide further insight into the strength of the variables as they relate to prosody ratings.

Table 17

Comparisons of Correlations with Data Sets

	4-4 Data Set		4-5 Data Set		5-5 Data Set	
	<i>t</i>	<i>df</i>	<i>t</i>	<i>df</i>	<i>t</i>	<i>df</i>
r_{12} vs. r_{13}	.44380	35	-2.04081*	35	-1.69310	34
r_{12} vs. r_{14}	5.07623**	35	5.38005**	35	4.59116**	34
r_{13} vs. r_{14}	-10.90539**	35	-5.41401**	35	-8.85857**	34

Note. r_{12} = Correlation between prosody-recoded and oral reading rate; r_{13} = Correlation between prosody-recoded and number of pauses; r_{14} = Correlation between prosody-recoded and oral reading accuracy.

Note. * $p < 0.05$. ** $p < 0.01$.

Both RATE and number of PAUSES share a strong association with PROSODY-RECODED membership. The correlations between PROSODY-RECODED and RATE and PROSODY-RECODED and PAUSES are significantly stronger than the correlation

between PROSODY-RECODED and ACCURACY for all data sets. In addition, the correlation between PROSODY-RECODED and RATE is not significantly different from the correlation between PROSODY-RECODED and PAUSES in the 4-4 data set or the 5-5 data set.

Summary of Results

The present study investigated the assessment of reading fluency through the measurement of reading prosody and the relation of prosodic measures to reading rate and other constructs of reading ability. Rating scales are used to evaluate prosodic reading, so the study explored the usefulness of this method. Because rating scales have been criticized for various reasons, including subjectivity, the study also sought to establish a more objective means of assessing fluency and prosody. Accounting for rate, pauses, and accuracy in oral reading were selected as objective measures of fluent and prosodic reading. In fact, two of these measures (rate and pauses) are reflected directly in two subscales (phrasing and pacing) from the Multidimensional Fluency Scale (Rasinski, 2004; Zutell & Rasinski, 1991) and the other (accuracy) is indirectly associated with a third subscale (smoothness).

Assessing prosody through the use of a rating scale produced distinct groups of readers. Descriptive profiles of these PROSODY groups were developed by examining online reading behaviors, and the groups clearly differed on these measures. In addition to measuring reading fluency through the use of PROSODY rating scales, cluster analysis was used to classify readers into FLUENCY groups based on more objective measures: (a) RATE, (b) PAUSES, and (c) ACCURACY. These analyses created three distinct

FLUENCY groups for each data set, 4-4, 4-5, and 5-5. Results were consistent across the three data sets.

Discriminant function analyses revealed that RATE and PAUSES were much better predictors of FLUENCY group membership than ACCURACY scores. In fact, RATE accounted for the most variance in all the analyses, followed by PAUSES. ACCURACY contributed much less to the FLUENCY groupings.

The comparisons between PROSODY rating scores and FLUENCY groupings demonstrated a high degree of agreement, though not perfect. In addition, the consistency of the three-group solution of FLUENCY for each data set led to a recoding of the original 5-level PROSODY ratings into three PROSODY-RECODED groups. These PROSODY-RECODED groups were shown to differ significantly from one another on RATE, PAUSES and WR-t. The low group was different from the middle and high groups on all three variables. The middle and high groups differed on RATE for all three data sets but differed on PAUSES only for the 4-4 and 5-5 data sets.

Correlations between the PROSODY-RECODED variable and oral reading variables (RATE, number of PAUSES, and ACCURACY) revealed the strength of RATE and PAUSES in characterizing groups (PROSODY-RECODED) of readers formed through the use of rating scales. RATE and PAUSES, in particular, are highly correlated with PROSODY-RECODED groups; in fact, they provide much of the information derived from rating scales.

Finally, statistical analyses revealed that the correlations between PROSODY-RECODED and RATE and between PROSODY-RECODED and PAUSES are significantly stronger than the correlation between PROSODY-RECODED and

ACCURACY for all three data sets. This indicates the relatively weak contribution of ACCURACY in characterizing groups formed through rating scales (PROSODY-RECODED).

Results from this study suggest that rating scales can be used accurately and productively in measuring young readers' fluency and prosody. However, the cluster analyses suggest that rating scales are most robust when used to distinguish three levels (low, middle, and high) of student performance. In addition, the data reveal that online measures of oral reading RATE, number of PAUSES, and to a lesser degree ACCURACY provide objective measures of fluency and prosody. These measures are less reliant on subjective interpretation and are easy to collect, especially reading RATE, which proved to be the most powerful predictor of fluency in these analyses.

Chapter Four: Discussion and Implications

Fluent reading is an important and well-acknowledged dimension of reading ability (Adams, 1990; Breznitz, 2006; NICHD, 2000; Perfetti, 1985). Despite the general recognition of its significance, however, how to define and assess fluency remains in question (Hudson et al., 2009; Kuhn et al., 2010). For example, some authorities argue that accuracy and rate of reading alone account for and define fluent reading (Carver, 1990; LaBerge & Samuels, 1974; Logan, 1988; Wolf & Katzir-Cohen, 2001). Others acknowledge accuracy and rate as important components of fluency but insist that prosodic form also plays a necessary role (Kuhn et al., 2010; Rasinski, 2003; Schreiber, 1987, 1991). Adding to the complexity of the issue is a divergence between researchers who argue that to be fluent a reader must not only exhibit prosodic form but also adequate expression (Allington, 1983; Klauda & Guthrie, 2008; Rasinski, 2003;). That is, fluent oral reading must represent the meaning and emotion of a text through the expressive use of the voice.

Acknowledging that natural prosodic reading is a marker of skilled reading raises the problem of assessment. Some have argued that fluency is measured best through the use of prosodic rating scales (Benjamin et al., 2013; Pinnell et al., 1995; Rasinski et al., 2011; Zutell & Rasinski, 1991). Others have countered that these are subjective and require training that is often unavailable to users, especially classroom teachers (Schwanenflugel et al., 2004), and recommend using objective measures like parsing techniques (Clay & Imlach, 1971; Kleiman et al., 1979; Koriat et al., 2002; Young &

Bowers, 1995) or even spectrographic analysis to establish prosodic features (Benjamin et al., 2013; Cowie et al., 2002; Dowhower, 1987; Kowal et al., 1985; Miller & Schwanenflugel, 2006, 2008; Schwanenflugel et al., 2004). These approaches, while objective, are not efficient, and in the case of spectrographic use are neither logistically nor economically feasible for a classroom teacher.

This study sought to address some of these issues by establishing a more efficient and objective measure of prosody and fluency. Groups formed by subjective judgments (prosody rating scales) of reading performance were compared to groups formed by more objective measures (rate, number of pauses, and accuracy) of reading performance. On the one hand, the validity of using a prosody rating scale is tested, and on the other the relations between prosody and fluency measures are examined. In addition, the study examines the contributions of rate, number of pauses, and accuracy toward a measurement of fluency.

Major Findings of the Study

Data from this study show that it is possible to listen to children read orally and judge with some agreement and confidence the prosody quality of the reading. Furthermore, the prosody scale ratings correlate to other objective measures of reading fluency performance, adding support to researchers who have made the case that one can listen to a reader and judge the quality of the reading (Benjamin et al., 2013; Cowie et al., 2002).

That being said, data from this study challenge the idea that fluent reading must involve expressiveness, the insistence that fluent oral reading displays the emotion and tone of a text through the expressive use of voice. This type of reading is a specific and

practiced skill, such as is attained by readers who offer an oral interpretation of a text for listeners, and this kind of performance is not necessary for ordinary skillful reading. In fact, this study found that natural prosodic reading—reading that articulated the grammatical and syntactic structure of a text through pauses, appropriate phrasing and pitch changes—characterized the most skillful readers. Yet expressive reading as defined above was not a part of their oral rendering, lending further support to Cowie et al.’s (2002) findings that fluency *permits* expressive reading but one *can read fluently without expressiveness*.

Results from this study also raise concern about the appropriateness of the widely used NAEP scale to accurately represent the oral reading behaviors of fourth and fifth graders. In addition, results seem to indicate that rating scales have limited discrimination ability—three levels: A teacher can identify readers below-, on-, and above-grade level fluency with confidence using a fluency scale, but ascribing reading differences beyond these levels is difficult, less accurate, time consuming, and impractical.

Most importantly, the argument that reading fluency must be measured by capturing reading prosody (Benjamin et al., 2013; Kuhn et al., 2010; Rasinski et al., 2011) through a rating scale is challenged by the results of this study. Objective measures of reading rate, number of pauses, and reading accuracy were used to form fluency groups, and all three measures predicted group membership. However reading rate proved to be the strongest and most consistent predictor of fluency group membership, reading accuracy the weakest. Based on the outcomes of the separate prosodic rating and objective fluency measures, students were placed in high, middle, and low fluency groups. These groupings by prosody ratings and by objective measures of reading

behavior were found to have a high degree of overlap. Analysis revealed that both pause and rate measures were highly correlated with a carefully used prosodic rating scale. That is, both captured objectively what the rating scale provided subjectively.

In addition, differences were observed in the fluency and prosodic control of students depending on the difficulty level of the text. That is, fourth-grade students read fourth grade passages more fluently than fifth grade passages. Those same students read fifth-grade passages as fifth-graders with a similar level of fluency as they had read fourth-grade passages as fourth graders. In other words, the difficulty of a passage has an impact not just on the accuracy of the reading but also the rate and prosodic form of the reading.

Implications for Teaching and Research

To use a prosodic rating scale effectively requires training and practice as well as sufficient time for repeated listening to a reading sample. In addition, judgments must be made about the quality of the reading, and these can be subject to doubt and uncertainty. It also should be noted that not all reading scales are the same. The research team began with the NAEP scale but abandoned it because it was not capturing the reading behavior of the participants. The rating scale that was used in this study came from the data, and that is a strength of the scale for this particular data set. But, another data set may require a slightly different scale. These issues demonstrate the subjective nature of rating scales. The Multidimensional Fluency Scale (Raskinski, 2003; Zutell & Rasinski, 1991) and perhaps the CORFS (Benjamin et al., 2013) are better choices for fluency rating scales for researchers interested in capturing subtle differences in reading fluency and prosody. By design, they are complex and time consuming to administer. However, for teachers

this level of analysis is impractical. Generally, a teacher needs only to discriminate readers in trouble with fluency from those who are not. In many cases, this can be determined without a fluency scale, just by listening to a reader, but if a scale is used it should be simple and transparent.

Marking pauses, an alternative to fluency rating scales, has proven to capture important components to fluent reading (Clay & Imlach, 1971; Cowie et al., 2002; Kowal et al., 1985; Young & Bowers, 1995) and did so in this study. Findings from this study show the importance of counting pauses in fluency research, along with collecting reading rate, accuracy, and ratings of prosody. However, marking pauses, though a strong objective measure of reading fluency, also requires a significant commitment of time. In other words, both rating scales and pause marking present practical challenges for teachers. Measuring reading rate, on the other hand, is a useful and objective alternative and requires little training. The strong correlations found between reading rate and the prosodic rating scale, number of pauses, and reading accuracy mean that teachers can employ rate measures to determine reading fluency among their students. These data support the use of reading rate as a proxy for reading fluency.

Reading rate is not an end in itself; rather it is a component of an overall picture of the reading process. Assessing rate along with accuracy (while students are reading for meaning) can provide important qualifying information. In particular, it can help describe the limits of an instructional level where accuracy alone does not reliably predict reading fluency. For example, a student could be found to read with 94% accuracy and adequate comprehension in a fifth-grade-level text but at a rate of 56 wpm. This would indicate severe problems with fluency and would make completing fifth-grade-level reading a

frustrating and unproductive labor. At a lower difficulty level, perhaps fourth grade, that student may read with sufficient fluency to provide a level of challenge suitable for improvements in reading ability.

Reading rate is of particular concern in North Carolina's current educational environment. The state's legislative requirement that a specific commercially marketed assessment be used in North Carolina schools brings this issue to the fore. Because Wireless Generation's mCLASS: Reading 3D uses extremely low accuracy scores in setting instructional levels (as low as 90%), and because Reading 3D ignores reading rate in these decisions, there exists a high potential for placing struggling readers in material too difficult for success. Paying attention to rate especially and the use of more appropriate accuracy cut offs (95%) will set instructional reading levels more accurately. These adjustments can make a significant difference for lower functioning readers.

Educational leaders should be concerned that the assessment system that is being required in our public schools in North Carolina has little research backing. Data from this study challenge the validity of Reading 3D; however, more research is needed. Reading rate has long been recognized by psychologists as being an important indicator of reading ability and fluency (Carver, 1990; Perfetti, 1985). Using reading rate in this way makes sense. What doesn't make sense is *not* using reading rate as a variable to determine children's reading levels. This is an important area for educational policy and research.

As researchers address these questions, data reduction techniques, such as cluster analysis used in this study, and factor analysis bring more rigor and objectivity to assessment questions. Informal Reading Inventories have been used for 70 years, but few

have had their measurement properties examined. In truth, most of these instruments are anything but informal, yet researchers need to take the time to document how these instruments work and how they relate to other instruments that are used to measure reading behavior. This is particularly important in the climate of public schools today with its heavy emphasis on assessment of students and teachers.

Limitations

A particular strength of this study is that it points to a close relationship between reading rate and fluent, prosodic reading. This offers teachers a practical and objective means for monitoring this dimension of reading and for making instructional adjustments when needed. Based on clinical experience, the expectation is that the findings would generalize, but studies with larger data sets need to confirm this.

This study limited its focus to fourth- and fifth-grade readers. Studying readers at these grades makes sense in that they are elementary school years when it has become essential that accurate fluent reading be well established. Further study, however, is needed to confirm that objectively and subjectively rated groups match at other grade levels. Also, participants for this study were drawn from predominantly rural areas. Therefore one might question whether the results reported here would extend, for example, to urban populations.

Because the difficulty level of passages read in this study was found to have an impact on fluency and prosodic form as well as accuracy, one must question the use of grade-level-only passages on the state's End of Grade (EOG) tests. If text difficulty has an impact on overall reading performance, then how is one to understand lower student performances? Is the information derived from such results in any way informative?

Could important information be gained by testing students on passages that are graded in difficulty? And too, careful research should be done on the purported difficulty levels of EOG reading passages. Do they accurately represent the reading levels to which they refer?

References

- Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: The MIT Press.
- Allington, R. L. (1983). Fluency: The neglected reading goal. *The Reading Teacher*, 36(6), 556-561.
- Arnold, J. M., & Sableski, M. K. (2006). Charles Hubbard Judd (1873-1946): A leader in silent and oral reading instruction. In S. E. Israel & E. J. Monaghan (Eds.), *Shaping the reading field: The impact of early reading pioneers, scientific research, and progressive ideas* (pp. 101-118). Newark, DE: International Reading Association.
- Bailey, A. L. & Drummond, K. V. (2006). Who is at risk and why? Teachers' reasons for concern and their understanding and assessment of early literacy. *Educational Assessment*, 11(3), 149-178.
- Benjamin, R. G., & Schwanenflugel, P. J. (2010). Text complexity and oral reading prosody in young readers. *Reading Research Quarterly*, 45(4), 388-404.
- Benjamin, R. G., Schwanenflugel, P. J., Meisinger, E. B., Groff, C., Kuhn, M. R., & Steiner, L. (2013). A spectrographically grounded scale for evaluating reading expressiveness. *Reading Research Quarterly*, 48(2), 105-133.

- Beswick, J. F., Willms, J. D., & Sloat, E. A. (2005). A comparative study of teacher ratings of emergent literacy skills and student performance on a standardized measure. *Education, 126*(1), 116-137.
- Breznitz, Z. (2006). *Fluency in reading: Synchronization of processes*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall.
- Carver, R. P. (1990). *Reading rate: A review of research and theory*. New York: Academic Press.
- Chall, J. S. (1983). *Stages of reading development*. New York: McGraw-Hill Book Company.
- Clay, M. M., & Imlach, R. H. (1971). Juncture, pitch, and stress as reading behavior variables. *Journal of Verbal Learning and Verbal Behavior, 10*, 133-139.
- Cobb, L. (1835). *The North American reader*. New York: B. & S. Collins.
- 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.
- Cromer, W. (1970). The difference model: A new explanation for some reading difficulties. *Journal of Educational Psychology, 61*(6), 471-483.

- Cross, D. R., & Paris, S. G. (1988). Developmental and instructional analyses of children's metacognition and reading comprehension. *Journal of Educational Psychology, 80*(2), 131-142.
- DeFord, D. E. (1991). Fluency in initial reading instruction: A reading recovery lesson. *Theory into Practice, 30*(3), 201-210.
- Dowhower, S. L. (1987). Effects of repeated reading on second-grade transitional readers' fluency and comprehension. *Reading Research Quarterly, 22*(4), 389-406.
- Dowhower, S. L. (1991). Speaking of prosody: Fluency's unattended bedfellow. *Theory into Practice, 30*(3), 165-175.
- Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185.
- Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Hoboken, NJ: Wiley.
- Gough, P. B., Hoover, W. A., & Peterson, C. L. (1996). Some observations on a simple view of reading. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 1-13). Mahwah, NJ: Lawrence Erlbaum.
- Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *RASE: Remedial and Special Education, 7*(1), 6-10.

- Hammett, L. A., Van Kleeck, A., & Huberty, C. J. (2003). Patterns of parents' extratextual interactions during book sharing with preschool children: A cluster analysis study. *Reading Research Quarterly, 38*(4), 442-468.
- Harris, A. & Jacobson, M. (1982). *Basic reading vocabularies*. New York: Macmillan.
- Henderson, E. H. (1981). *Learning to read and spell: The child's knowledge of words*. DeKalb, IL: Northern Illinois University Press.
- Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2*(2), 127-160.
- Hudson, R. F., Lane, H. B., & Pullen, P. C. (2005). Reading fluency assessment and instruction: What, why, and how? *The Reading Teacher, 58*(8), 702-714.
- Hudson, R. F., Pullen, P. C., Lane, H. B., & Torgesen, J. K. (2009). The complex nature of reading fluency: A multidimensional view. *Reading & Writing Quarterly, 25*, 4-32.
- Huey, E. B. (1908/1968). *The psychology and pedagogy of reading*. Boston: MIT Press.
- Judd, C. H. (1921). Analysis of learning processes and specific teaching. *Elementary School Journal, 21*(9), 655-664.
- Klauda, S. L., & Guthrie, J. T. (2008). Relationships of three components of reading fluency to reading comprehension. *Journal of Educational Psychology, 100*(2), 310-321.

- Kleiman, G. M., Winograd, P. N., & Humphrey, M. M. (1979). *Prosody and children's parsing of sentences*. (Technical Report No. 123). Urbana-Champaign, IL: Center for the Study of Reading, University of Illinois.
- Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275-304.
- Kojic-Sabo, I., & Lightbown, P. M. (1999). Students' approaches to vocabulary learning and their relationship to success. *The Modern Language Journal, 83*(2), 176-192.
- Koriat, A., Greenberg, S. N., & Kreiner, H. (2002). The extraction of structure during reading: Evidence from reading prosody. *Memory & Cognition, 30*(2), 270-280.
- Kowal, S., O'Connell, D. C., O'Brien, E. A., & Bryant, E. T. (1975). Temporal aspects of reading aloud and speaking: Three experiments. *The American Journal of Psychology, 88*(4), 549-569.
- Kuhn, M. R., Schwanenflugel, P. J., & Meisinger, E. B. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly, 45*(2), 230-251.
- Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology, 95*(1), 3-21.
- LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*(2), 293-323.
- Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*(4), 492-527.

- MacGinitie, W. H., & MacGinitie, R. K. (1989). *Gates-MacGinitie reading tests*. Rolling Meadows, IL: Riverside Publishing Company.
- Mann, H. (1867). *Lectures and annual reports on education*. Cambridge, MA.
- Meyer, M. S., & Felton, R. H. (1999). Repeated reading to enhance fluency: Old approaches and new directions. *Annals of Dyslexia, 49*, 283-306.
- Micro Power & Light Company. (2008). *Readability calculations for Windows*. Dallas, TX: Author.
- 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.
- 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.
- Morris, D. (2008). *Diagnosis and correction of reading problems*. New York: The Guilford Press.
- Morris, D., Bloodgood, J. W., Perney, J., Frye, E. M., Kucan, L., Trathen, W., Ward, D., & Schlagal, R. (2011). Validating craft knowledge: An empirical examination of elementary-grade students' performance on an informal reading assessment. *Elementary School Journal, 112*(2), 205-233.

- Morris, D., Trathen, W., Frye, E. M., Kucan, L., Ward, D., Schlagal, R., & Hendrix, M. (2013). The role of reading rate in the informal assessment of reading ability. *Literacy Research and Instruction, 52*(1), 52-64.
- Morris, D., Trathen, W., Lomax, R. G., Perney, J., Kucan, L., Frye, E. M., . . . Schlagal, R. (2012). Modeling aspects of print-processing skill: Implications for reading assessment. *Reading and Writing: An Interdisciplinary Journal, 25*(1), 189-215.
- National Institute of Child Health and Human Development (NICHD). (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.
- Norusis, M. J., & SPSS, Inc. (2010). *PASW Statistics 18 advanced statistical procedures*. Upper Saddle River, NJ: Prentice Hall.
- Park, H. M. (2008). *Univariate analysis and normality test using SAS, STATA, and SPSS. Working Paper*. The University Information Technology Services (UITs), Statistical and Mathematical Computing, Indiana University.

<http://indiana.edu/~statmath/stat/all/normality/normality.pdf>
- Parker, S. C. (1921). How to teach beginning reading. *The Elementary School Journal, 22*(4), 255-268.

- Pearson, P. D. (2002). American reading instruction since 1967. In N. B. Smith (Special Ed.), *American reading instruction* (special edition) (pp. 419-486). Newark, DE: International Reading Association.
- Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.
- Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*(4), 357-383.
- Pikulski, J. (2006). Fluency: A developmental and language perspective. In S. J. Samuels & A. Farstrup (Eds.), *What research has to say about fluency instruction* (pp. 70-93). Newark, DE: International Reading Association.
- Pikulski, J. J., & Chard, D. J. (2005). Fluency: Bridge between decoding and reading comprehension. *The Reading Teacher, 58*(6), 510-519.
- 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*. Washington, DC: National Center for Education Statistics, U.S. Department of Education.
- Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English Language* (2nd ed.). New York: Longman.
- Rasinski, T. V. (2003). *The fluent reader*. New York: Scholastic Professional Books.
- Rasinski, T. V. (2004). Creating fluent readers. *Educational Leadership, 61*(6), 46-51.

- Rasinski, T. V. (2006a). A brief history of reading fluency. In S. J. Samuels & A. E. Farstrup (Eds.), *What research has to say about fluency instruction* (pp. 4-23). Newark, DE: International Reading Association.
- Rasinski, T. V. (2006b). Reading fluency instruction: Moving beyond accuracy, automaticity, and prosody. *The Reading Teacher*, 59(7), 704-706.
- Rasinski, T. V., Reutzel, D. R., Chard, D., & Linan-Thompson, S. (2011). Reading fluency. In M. L. Kamil, P. D. Pearson, E. B. Moje, & P. P. Afflerbach (Eds.), *Handbook of reading research, IV* (pp. 286-319). New York: Routledge.
- Rasinski, T. V., Rikli, A., & Johnston, S. (2009). Reading fluency: More than automaticity? More than a concern for primary grades? *Literacy Research and Instruction*, 48(4), 350-361.
- Rayner, K., & Pollatsek, A. (1989). *The psychology of reading*. New York: Prentice-Hall.
- Samuels, S. J. (1988). Decoding and automaticity: Helping poor readers become automatic at word recognition. *The Reading Teacher*, 41, 756-760.
- Samuels, S. J. (2006). Looking backward: Reflections on a career in reading. *Journal of Literacy Research*, 38(3), 327-344.
- Samuels, S. J., Schermer, N., & Reinking, D. (1992). Reading fluency: Techniques for making decoding automatic. In S. J. Samuels & A. E. Farstrup (Eds.), *What research has to say about reading instruction* (pp. 124-144). Newark, DE: International Reading Association.

- Schatschneider, C., & Lonigan, C. (2010). Misunderstood statistical assumptions undermine criticism of the National Early Literacy Panel's Report. *Educational Researchers, 39*(4), 347-351.
- Schreiber, P. A. (1980). On the acquisition of reading fluency. *Journal of Reading Behavior, 12*(3), 177-186.
- 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). New York: Academic Press.
- Schreiber, P. A. (1991). Understanding prosody's role in reading acquisition. *Theory into Practice, 30*(3), 158-164.
- Schreiber, P. A., & Read, C. (1980). Children's use of phonetic cues in spelling, parsing, and – maybe – reading. *Bulletin of the Orton Society, 30*, 209-224.
- Schwanenflugel, P. J., Hamilton, A. M., Kuhn, M. R., Wisenbaker, J. M., & Stahl, S. A. (2004). Becoming a fluent reader: Reading skill and prosodic features in the oral reading of young readers. *Journal of Educational Psychology, 96*(1), 119-129.
- Share, D. L. (1995). Phonological recoding and self-teaching: *Sine qua non* of reading acquisition. *Cognition, 55*, 151-218.
- Smith, N. B. (2002). *American reading instruction* (Special ed.), Newark, DE: International Reading Association.
- Spear-Swerling, L., & Sternberg, R. J. (1996). *Off-track: When poor readers become "learning disabled."* Boulder, CO: Westview Press.

- Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Strecker, S. K., Roser, N. L., & Martinez, M. G. (1998). Toward understanding oral reading fluency. In T. Shanahan & F. V. Rodriguez-Brown (Eds.), *47th yearbook of the National Reading Conference* (pp. 295-310). Chicago: National Reading Conference.
- Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper Collins.
- Torgesen, J. K., & Hudson, R. F. (2006). Reading fluency: Critical issues for struggling readers. In S. J. Samuels, & A. E. Farstrup (Eds.), *What research has to say about fluency instruction* (pp. 130-158). Newark, DE: International Reading Association.
- Torgesen, J. K., Wagner, R. K., & Rashotte, C. (1999). *Test of word reading efficiency*. Austin, TX: Pro-Ed.
- Wade, S. E., Trathen, W., & Schraw, G. (1990). An analysis of spontaneous study strategies. *Reading Research Quarterly*, 25(2), 147-166.
- Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.
- White, S. (1995). *Listening to children read aloud: Oral fluency*. Washington, D. C.: National Center for Education Statistics.
- Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Scientific Studies of Reading*, 5(3), 211-239.

- Young, A., & Bowers, P. G. (1995). Individual differences and text difficulty determinants of reading fluency and expressiveness. *Journal of Experimental Child Psychology*, 60, 428-454.
- Zeno, S., Ivens, S. Millard, R., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science Associates.
- Zutell, J., & Rasinski, T. V. (1991). Training teachers to attend to their students' oral reading fluency. *Theory into Practice*, 30(3), 211-217.

Appendix A
Descriptive Statistics for the Print-Processing and Comprehension Measures
(Second-Sixth Grade)

Grade	Word Rec -timed (%)	Word Rec -untimed (%)	Oral Reading Accuracy (%)	Spelling (%)	Oral Rate (wpm)	Silent Rate (wpm)	Oral Reading Compreh. (%)	Silent Reading Compreh. (%)
Second (<i>n</i> = 134)	85 (19)	96 (11)	95 (4.8)	64 (23)	107 (38)	130 (52)	87 (17)	83 (20)
Third (<i>n</i> = 260)	82 (20)	95 (11)	96 (3.9)	65 (25)	119 (36)	150 (55)	85 (17)	80 (24)
Fourth (<i>n</i> = 238)	84 (19)	94 (11)	96 (3.6)	67 (26)	127 (34)	166 (60)	82 (18)	77 (22)
Fifth (<i>n</i> = 220)	83 (18)	94 (10)	96 (3.0)	65 (28)	128 (34)	172 (61)	81 (21)	76 (23)
Sixth (<i>n</i> = 106)	80 (18)	94 (12)	96 (2.7)	49 (27)	128 (35)	171 (52)	85 (20)	81 (21)

Morris, Bloodgood, Perney, Frye, Kucan, Trathen, Ward, & Schlagal (2011).

Appendix B
Word Recognition Measure

Word Recognition Lists (Preprimer through Eighth Grade)

<u>Preprimer</u>	<u>Primer</u>	<u>First</u>	<u>Second</u>	<u>Third</u>
and	back	leg	able	accept
cat	eat	black	break	favor
me	sun	smile	pull	seal
is	bird	hurt	week	buffalo
go	pat	dark	gate	slipper
play	saw	white	felt	receive
where	feet	couldn't	north	legend
like	lake	seen	rush	haircut
thing	hid	until	wrote	dresser
old	cut	because	perfect	icy
your	about	men	change	customer
up	one	winter	basket	thread
said	rain	shout	shoot	plop
big	water	glass	hospital	bandage
for	two	paint	spill	further
by	how	children	dug	moat
dog	window	table	crayon	closet
not	need	stand	third	unroll
who	that's	head	taken	storyteller
here	mother	drove	prize	yarn

Morris, Bloodgood, Perney, Frye, Kucan, Trathen, Ward, & Schlagal (2011).

Word Recognition Lists (Preprimer through Eighth Grade)

<u>Fourth</u>	<u>Fifth</u>	<u>Sixth</u>	<u>Seventh</u>	<u>Eighth</u>
average	labor	elevate	civic	administration
hamster	cripple	conservation	shirttail	federation
select	hasten	tenderness	nominated	militia
tobacco	frontier	barrier	gruesome	shambles
brilliant	riverbed	adulthood	disadvantage	bankrupt
liberty	settlement	kennel	architecture	goldenrod
prance	absent	humiliated	tonic	perishable
solemn	dissolve	nonfiction	straightforward	toddler
disease	plea	revive	warrant	cavernous
impress	surrender	wallet	unthinkable	imperative
miracle	organization	depression	ridicule	notorious
wrestle	evidence	carvings	engulf	subconscious
coward	width	similarity	kindhearted	corps
explode	rampaging	unanswered	maturity	laborious
opinion	horseshoe	fangernail	impassable	rivet
suffer	grammar	breed	bolster	unimaginable
vast	assorted	marrow	copyright	dizzily
relationship	soybean	starter	foliage	irritability
furnace	troublesome	pedestrian	prune	puncture
clan	circumstance	quantity	persecution	wholehearted

Morris, Bloodgood, Perney, Frye, Kucan, Trathen, Ward, & Schlagal (2011).

Internal Consistency, Stability and Hierarchical Difficulty of Word Recognition Measure

KR-21 coefficients provided an estimate of the internal consistency of the word recognition scores. KR-21 coefficients for WR-t ($k = 20$) across grades 2 to 6 ranged from .81 -- .87 (median = .86); coefficients for WR-ut ($k = 20$) ranged from .74 -- .86 (median = .85). The researchers also determined the stability of the word recognition scores from year to year (e.g., second to third, third to fourth, and so on). Stability coefficients for WR-t across grades 2 to 6 ranged from .83 -- .89 (median = .86); coefficients for WR-ut ranged from .78 -- .93 (median = .83).

For both WR-t and WR-ut, the word recognition lists proved to be hierarchical in difficulty. That is, at each grade level (2 to 6), the students read the grade-level list more accurately than they read the one-grade-level-above list (WR-t mean difference = 13% [range = 7% to 25%]; WR-ut mean difference = 5% [range = 2% to 10%]).
Morris, Bloodgood, Perney, Frye, Kucan, Trathen, Ward, & Schlagal (2011).

Appendix C
Passage Reading Inventory (Forms A, B, C, and D)

Grade Level	Form A	Form B	Form C	Form D
4	Woods & Moe (2003) (Form A) Crossing the River”	Leslie & Caldwell (2006) (Level 4) “Johnny Applesseed”	Woods & Moe (2003) (Form C) “The Beloved Horse”	Leslie & Caldwell (2006) (Level 4) “Amelia Earhart”
5	Leslie & Caldwell (1990) (Level 5) “Christopher Columbus”	Woods & Moe (2003) (Form A) “The Bicycle Race”	Leslie & Caldwell (2006) (Level 5) “Martin Luther King”	Roe & Burns (2007) (Form A) “Harbor Seal Pup”

Morris, Bloodgood, Perney, Frye, Kucan, Trathen, Ward, & Schlagal (2011).

Appendix D

Passage Readability, Equivalency, Hierarchical Difficulty, and Stability

The readability of each passage was calculated, using the New Dale-Chall and Fry formulas (Micro Power & Light Company, 2008) for fourth and fifth grade. The Dale-Chall formula, which emphasizes vocabulary difficulty, and the Fry formula, which emphasizes orthographic complexity (number of syllables), rated passage readability separately, but the two measures were combined to determine approximate grade level readability. Dale-Chall median readabilities were *upper-third grade* for the fourth-grade passages and *mid-fourth grade* for the fifth-grade passages. The corresponding Fry median readabilities were higher: *Fourth grade* for the fourth-grade passages and *sixth-grade* for the fifth-grade passages. The average of the Dale-Chall and Fry scores, at each grade, approximates grade-level readability.

Student scores on the passages were compared to determine if the different forms of the passages were equivalent at grade levels. One-way ANOVAs at each grade level showed that, for oral reading accuracy (ORA) and oral reading rate (ORR), the four passage-reading forms (A, B, C, and D) were roughly equivalent; that is, no statistically significant differences in performance ($p > .01$) were observed. For ORA and ORR, the passages also proved to be hierarchical in difficulty. That is, at each grade level, the students read the grade-level passage more accurately and more quickly than they read the one-grade-level-above passage (ORA mean difference = 1.4% [range = .4% to 2.0%]; ORR mean difference = 12 wpm [range = 3 wpm to 16 wpm]).

Regarding the consistency of the oral reading measures from one year to the next, stability coefficients for ORA across grades ranged from .81 -- .86 (median = .83);

coefficients for ORR ranged from .84 -- .94 (median = .91). Thus, the data are stable across grades and years.

Morris, Bloodgood, Perney, Frye, Kucan, Trathen, Ward, & Schlagal (2011).

Appendix E Screening Data

Prior to analyses, data sets were screened for missing values and non-normality. A discussion of the methods undertaken to deal with missing values and non-normal distributions is presented below.

Missing Data

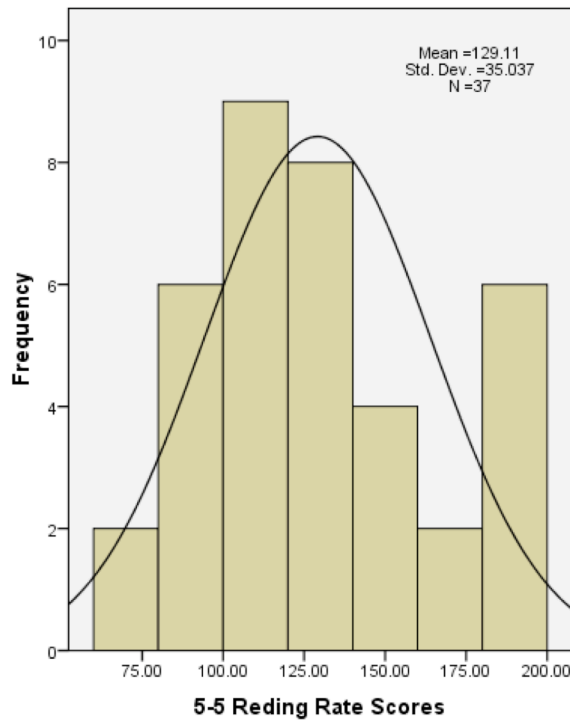
Data for this study include 38 subjects, each with three data sets, resulting in 114 potential data records for each of the five variables for a total of 570 data records. Only two data records from the total were replaced; replacement data accounted for less than 0.3% of the entire data set. Before analyses, the data were screened for missing values. In two cases (Cases 2 and 3) in the 4-5 data set, students did not complete all the sections of the reading assessment that were required in the present study. The missing data were two scores for timed word recognition (less than 2% of WR-t scores) that needed to be imputed. Group means were used to adjust the missing data for a student. The mean difference from one year to the next was calculated for timed word recognition, and this amount was subtracted from the student's last recorded score for the variable in question. Using the mean difference is a conservative method of imputing missing data.

Distributional Normality of Predictor and Dependent Variables

Graphical and numerical methods were conducted to test normality of the distributions of scores on the predictor and dependent variables. Summary statistics such as skewness and kurtosis were obtained from numerical methods, and statistical theory-driven tests of normality were conducted. Skewness is a measure of dispersion in the distribution. It measures the degree to which data values deviate from the mean to either

the left tail of the distribution (positive skew) or the right tail (negative distribution). A non-zero skew score is also an indication of the direction of the asymmetry; a positive skew score means the data are positively skewed while a negative score indicates the data are piled towards the right end of the distribution away from the mean. A zero score indicates no skew in the data set. Kurtosis, another dispersion measure, is a measure of the “peakedness” or flatness in the data relative to a normal distribution. Highly kurtotic data sets are characterized by a swarm of data peaked around the mean with short tails. On the other hand, a flat top and long tails characterize a data set with low kurtosis. Peaked distributions are associated with a positive kurtosis value whereas a negative kurtosis is associated with a relatively flat distribution (Brown, 1996).

In addition to numerical (i.e., skewness, kurtosis) and graphical (e.g., box plots, histograms) methods, which provided objective and intuitive ways of examining normality in the data respectively, the Shapiro-Wilk (W) statistic was used for testing normality. The W statistic is recommended (Park, 2008) for samples sizes greater than or equal to 7 and less than or equal to 2,000. The W is reported as a positive number, less than or equal to 1. A W score close to 1 indicates a normal distribution of data. For example, according to the Shapiro-Wilk test, the reading rate scores for the 5-5 data set are normally distributed, $W = .948$, $df = 37$, $p = .081$ (see Figure 7 below). In other words, the null hypothesis of normality is not rejected because the actual probability level ($p = .081$) is greater than the nominal probability level of .05.



The W was computed for all dependent and predictor variables used in the analyses. The W test results and other descriptive statistics are listed below for all variables. Data sets with significant ($p < .05$) W values are marked with an asterisk indicating that the group is not normally distributed. There were 10 distributions across all three data sets for which the W test was significant, indicating non-normality. For these “non-normal” data sets, standard errors of skewness (SES) and standard errors of kurtosis (SEK) were calculated to provide ranges of acceptable skewness and kurtosis values. According to Brown (1996), a skew value more than 2 X SES and a kurtosis value more than 2 X SEK are significant departures from acceptable values.

SES is estimated by taking the square root of 6 over sample size (n) (Tabachnick & Fidell, 1996). To use Brown’s example, if a data set derived from 30 subjects has a

skewness statistic of $-.9814$, is there significant deviation from normality? The *SES* of this data set is $.4472$. By multiplying the *SES* with 2, the range of acceptable skewness values can be obtained: for this data set $.8944$ constitutes the upper bound and $-.8944$ the lower bound. Since the absolute value of the skewness statistic is $-.9814$, which is greater than $-.8944$, the data set appears to be offended by significant departure from normality. We can't assume therefore that skewness observed in the data set can be due to chance fluctuations in the skewness statistic. A skewness value within the range of $.8944$ and $-.8944$, however, would have indicated a distribution with no significant skewness problem.

SEK is estimated by taking the square root of 24 over sample size (n) (Tabachnick & Fidell, 1996). For the same example, *SEK* is 1.7888 ($.8944 \times 2$) and the absolute value of the kurtosis statistic is 1.9142 , which is greater than 1.7888 . Because the kurtosis statistic of the data set is beyond the upper bound of the acceptable range, departure from normality is beyond chance fluctuations in the kurtosis statistic.

Of the 10 non-normal distributions, eight have skewness and kurtosis values within acceptable *SES* and *SEK* ranges. Their departure from normality could be explained by chance fluctuations in the skewness and kurtosis measures. Only two of 54 distributions have skewness and kurtosis values that are beyond acceptable ranges; the two distributions are the ACCURACY scores for the middle FLUENCY group on the 4-4 data set and the WR-t scores for the middle PROSODY-RECODED group on the 4-5 data set. The non-normality of both distributions most likely results from most of the scores approaching the ceiling with one outlier in each group.

Descriptive and Normality Statistics of Predictor and Dependent Variables

4-4 Data Set							
Predictor Variables in the Discriminant Function Analysis by Cluster-Formed FLUENCY Groups				Dependent Variables in Analyses of Variance (ANOVAs) by Prosody Scale-Formed PROSODY Groups			
	<i>Mean</i> (<i>SD</i>) [<i>range</i>]	<i>Skewness</i> (<i>SES</i>) <i>Kurtosis</i> (<i>SEK</i>)	Shapiro-Wilk Normality Test (<i>W</i>)		<i>Mean</i> (<i>SD</i>) [<i>range</i>]	<i>Skewness</i> (<i>SES</i>) <i>Kurtosis</i> (<i>SEK</i>)	Shapiro-Wilk Normality Test (<i>W</i>)
RATE				RATE			
High	169.21 (18.253) [143-218]	.926 1.260	<i>W</i> = .924 <i>df</i> = 19 <i>p</i> = .136	High	171.62 (18.913) [143-218]	.671 .983	<i>W</i> = .942 <i>df</i> = 16 <i>p</i> = .369
Middle	122 (13.702) [102-140]	-1.111 -1.408	<i>W</i> = .949 <i>df</i> = 9 <i>p</i> = .675	Middle	135.20 (17.812) [106-161]	-.086 -.859	<i>W</i> = .974 <i>df</i> = 10 <i>p</i> = .924
Low ^a	95.4 (12.98) [70-107]	-1.043 (1.549) -.167 (3.0983)	<i>W</i> = .844 <i>df</i> = 10 <i>p</i> = .049*	Low	97.41 (12.866) [70-113]	.404 -1.116	<i>W</i> = .875 <i>df</i> = 12 <i>p</i> = .076
PAUSES				PAUSES			
High	25.36 (3.386) [17-30]	-.796 .665	<i>W</i> = .946 <i>df</i> = 19 <i>p</i> = .335	High	24.75 (3.296) [17-30]	-.731 .776	<i>W</i> = .954 <i>df</i> = 16 <i>p</i> = .548
Middle	35.33 (4.663) [30-42]	.470 -1.242	<i>W</i> = .889 <i>df</i> = 9 <i>p</i> = .195	Middle	32.3 (4.243) [27-42]	1.310 2.331	<i>W</i> = .890 <i>df</i> = 10 <i>p</i> = .168
Low	47.5 (4.927) [39-55]	-.209 -.664	<i>W</i> = .963 <i>df</i> = 10 <i>p</i> = .824	Low	46.33 (5.262) [39-55]	.093 -1.130	<i>W</i> = .945 <i>df</i> = 12 <i>p</i> = .560
ACCURACY				WR-t			
High	96.94 (2.067) [93-100]	-.512 -.565	<i>W</i> = .925 <i>df</i> = 19 <i>p</i> = .143	High ^d	96.56 (3.966) [90-100]	-.662 (-1.2247) -1.006 (-2.4494)	<i>W</i> = .768 <i>df</i> = 16 <i>p</i> = .001*
Middle ^b	97.44 (2.242) [92-99]	-2.087 (1.6329) 4.989 (3.2659)	<i>W</i> = .713 <i>df</i> = 9 <i>p</i> = .002*	Middle	87 (8.881) [70-100]	-.464 .054	<i>W</i> = .965 <i>df</i> = 10 <i>p</i> = .841
Low ^c	92.9 (1.101) [92-95]	.863 (1.5491) -.522 (3.0983)	<i>W</i> = .810 <i>df</i> = 10 <i>p</i> = .019*	Low	69.16 (16.764) [40-95]	-.469 -.673	<i>W</i> = .933 <i>df</i> = 12 <i>p</i> = .413

Descriptive and Normality Statistics of Predictor and Dependent Variables

4-5 Data Set							
Predictor Variables in the Discriminant Function Analysis by Cluster-Formed FLUENCY Groups				Dependent Variables in Analyses of Variance (ANOVAs) by Prosody Scale-Formed PROSODY Groups			
	<i>Mean</i> (<i>SD</i>) [<i>range</i>]	<i>Skewness</i> (<i>SES</i>) <i>Kurtosis</i> (<i>SEK</i>)	Shapiro-Wilk Normality Test (<i>W</i>)		<i>Mean</i> (<i>SD</i>) [<i>range</i>]	<i>Skewness</i> (<i>SES</i>) <i>Kurtosis</i> (<i>SEK</i>)	Shapiro-Wilk Normality Test (<i>W</i>)
RATE				RATE			
High	147.8 (12.389) [129-168]	.168 -0.877	<i>W</i> = .977 <i>df</i> = 10 <i>p</i> = .946	High	153 (11.284) [135-168]	-.491 -0.308	<i>W</i> = .957 <i>df</i> = 7 <i>p</i> = .797
Middle	128.91 (15.397) [111-153]	.64 -1.090	<i>W</i> = .887 <i>df</i> = 11 <i>p</i> = .129	Middle	132.36 (12.387) [116-153]	.400 -1.110	<i>W</i> = .945 <i>df</i> = 11 <i>p</i> = .585
Low	89.58 (13.564) [67-116]	.423 -0.132	<i>W</i> = .957 <i>df</i> = 17 <i>p</i> = .576	Low	94.6 (18.517) [67-145]	.985 1.435	<i>W</i> = .930 <i>df</i> = 20 <i>p</i> = .156
PAUSES				PAUSES			
High	30.6 (3.272) [26-36]	.231 -1.070	<i>W</i> = .957 <i>df</i> = 10 <i>p</i> = .746	High	30.57 (3.644) [26-36]	.627 -0.800	<i>W</i> = .905 <i>df</i> = 7 <i>p</i> = .362
Middle	38.09 (2.981) [34-43]	.252 -1.094	<i>W</i> = .953 <i>df</i> = 11 <i>p</i> = .679	Middle	36.09 (4.826) [27-44]	-.048 .054	<i>W</i> = .959 <i>df</i> = 11 <i>p</i> = .754
Low	50.52 (7.706) [40-70]	1.011 1.265	<i>W</i> = .919 <i>df</i> = 17 <i>p</i> = .144	Low	48.65 (8.412) [37-70]	.834 .753	<i>W</i> = .937 <i>df</i> = 20 <i>p</i> = .210
ACCURACY				WR-t			
High	95 (1.885) [92-99]	.621 1.807	<i>W</i> = .922 <i>df</i> = 10 <i>p</i> = .374	High	90 (7.637) [75-100]	-1.178 3.000	<i>W</i> = .835 <i>df</i> = 7 <i>p</i> = .089
Middle ^c	97 (.894) [96-98]	.000 -1.875 (2.9541)	<i>W</i> = .795 <i>df</i> = 11 <i>p</i> = .008*	Middle ^f	90.91 (9.954) [65-100]	-1.978 (1.4770) 4.574 (2.9541)	<i>W</i> = .766 <i>df</i> = 11 <i>p</i> = .003*
Low	93.94 (2.014) [91-97]	.091 -1.302	<i>W</i> = .929 <i>df</i> = 17 <i>p</i> = .212	Low	61.25 (21.391) [30-95]	.145 -1.130	<i>W</i> = .938 <i>df</i> = 20 <i>p</i> = .221

Descriptive and Normality Statistics of Predictor and Dependent Variables

5-5 Data Set							
Predictor Variables in the Discriminant Function Analysis by Cluster-Formed FLUENCY Groups				Dependent Variables in Analyses of Variance (ANOVAs) by Prosody Scale-Formed PROSODY Groups			
	<i>Mean</i> (<i>SD</i>) [<i>range</i>]	<i>Skewness</i> (<i>SES</i>) <i>Kurtosis</i> (<i>SEK</i>)	Shapiro-Wilk Normality Test (<i>W</i>)		<i>Mean</i> (<i>SD</i>) [<i>range</i>]	<i>Skewness</i> (<i>SES</i>) <i>Kurtosis</i> (<i>SEK</i>)	Shapiro-Wilk Normality Test (<i>W</i>)
RATE				RATE			
High	182.87 (8.078) [171-197]	.376 .206	<i>W</i> = .974 <i>df</i> = 8 <i>p</i> = .929	High	173.36 (17.71) [145-197]	-.659 -.896	<i>W</i> = .890 <i>df</i> = 11 <i>p</i> = .141
Middle	132.66 (12.901) [115-152]	.025 -1.542	<i>W</i> = .928 <i>df</i> = 12 <i>p</i> = .359	Middle	125.31 (11.101) [107-145]	.257 -.841	<i>W</i> = .953 <i>df</i> = 13 <i>p</i> = .648
Low	101.29 (18.499) [72-137]	.156 -.967	<i>W</i> = .957 <i>df</i> = 17 <i>p</i> = .568	Low	95.46 (16.148) [72-122]	.355 -1.190	<i>W</i> = .937 <i>df</i> = 13 <i>p</i> = .422
PAUSES				PAUSES			
High	28.37 (4.657) [21-35]	-.170 -.684	<i>W</i> = .988 <i>df</i> = 8 <i>p</i> = .991	High	30 (5.31) [21-40]	.132 .195	<i>W</i> = .992 <i>df</i> = 11 <i>p</i> = .999
Middle ^h	35.33 (5.757) [29-44]	.486 (1.4142) -1.532 (2.8284)	<i>W</i> = .859 <i>df</i> = 12 <i>p</i> = .047*	Middle	38.84 (7.765) [29-52]	.248 -1.154	<i>W</i> = .922 <i>df</i> = 13 <i>p</i> = .266
Low	55.88 (10.03) [38-73]	.048 -.932	<i>W</i> = .971 <i>df</i> = 17 <i>p</i> = .835	Low	58.92 (9.031) [45-73]	-.127 -1.122	<i>W</i> = .952 <i>df</i> = 13 <i>p</i> = .625
ACCURACY				WR-t			
High	97.37 (1.685) [95-100]	.168 -.913	<i>W</i> = .966 <i>df</i> = 8 <i>p</i> = .862	High ^l	96.36 (3.233) [90-100]	-.291 (1.4770) -.208 (2.9541)	<i>W</i> = .793 <i>df</i> = 11 <i>p</i> = .008*
Middle	97.25 (1.138) [95-99]	-.583 -.138	<i>W</i> = .912 <i>df</i> = 12 <i>p</i> = .228	Middle	88.07 (8.548) [70-100]	-.942 .398	<i>W</i> = .897 <i>df</i> = 13 <i>p</i> = .122
Low	93.64 (2.498) [89-98]	-.225 -.644	<i>W</i> = .961 <i>df</i> = 17 <i>p</i> = .655	Low	72.307 (13.481) [40-90]	-1.156 1.663	<i>W</i> = .915 <i>df</i> = 13 <i>p</i> = .214

Vita

A resident of Mocksville, North Carolina, until her death in 2013, Mary Proctor Hendrix received her Bachelor of Science in Biology Education from Garner-Webb University in 1982 and her Master of Arts in Reading Education from Appalachian State University in 2002. Dr. Hendrix served as the Title 1 Reading Teacher at Cooleemee Elementary School in Davie County, North Carolina, from 2000 through 2007, as well as teaching part-time at Appalachian State University in the Department of Reading Education and Special Education.

Dr. Hendrix is co-author of two publications and four presentations on reading fluency and is a member of the International Reading Association, the Literacy Research Association, and the American Reading Forum. She has received a number of awards for her academic success, including induction into the Gallery of Distinguished Alumni at Gardner-Webb University, the Alpha Epsilon Lambda National Graduate Student Honor Society, and the Phi Kappa Phi Honor Society.

Dr. Hendrix was awarded the Doctor of Education in Educational Leadership from Appalachian State University posthumously in May 2013.