2012

Evaluating a Brief Measure of Reading Comprehension for Narrative and Expository Text: The Convergent and Predictive Validity of the Reading Retell Rubric

Lisa Beth Thomas
Lehigh University

Evaluating a Brief Measure of Reading Comprehension for Narrative and Expository Text: The Convergent and Predictive Validity of the Reading Retell Rubric

by

Lisa B. Thomas

Presented to the Graduate and Research Committee

of Lehigh University

In Candidacy for the Degree of Doctor of Philosophy

in

School Psychology

Lehigh University

April 20, 2012

Certificate of Approval

Approved and recommended for acceptance as a dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Date

Edward S. Shapiro, Ph.D.
Dissertation Chair
Professor of School Psychology
Lehigh University

Accepted Date

Committee Members:

Nanette S. Fritschmann, Ph.D.
Coordinator of Instructional Services
Orange County Department of Education
Department of Education Intern Council Member
University of California, Irvine

Robin Hojnoski, Ph.D.
Assistant Professor of School Psychology
Lehigh University

Minyi Shih, Ph.D.
Assistant Professor of Special Education
Lehigh University

Acknowledgements

There are many people I would like to thank who have supported me on this journey and made the completion of this dissertation possible. First, I must thank Dr. Craig Edelbrock for introducing me to the field of school psychology and encouraging me to pursue graduate studies at Lehigh University. Next, I would like to thank Dr. George DuPaul for inspiring me to become a school psychologist and affording me the opportunity to work as a graduate assistant on one of his research projects. Your support and guidance throughout the years has helped me grow as both a researcher and practitioner. I would also like to extend my sincerest appreciation and gratitude to the faculty and staff of Lehigh University's College of Education who have supported me throughout my graduate studies. I am especially thankful for the guidance and insight from my committee members, Dr. Nanette Fritschmann, Dr. Robin Hojnoski, and Dr. Minyi Shih. Their support and feedback were invaluable throughout the development and completion of my dissertation. In addition, I am indebted to the students and staff who participated in this study, without whom this study would not have been possible. I am also eternally grateful to Kristen Carson, Amanda Curry, Jennifer Parks, and Heather Mushock for their time and dedication to data collection and scoring. I would also like to thank Amanda Deanne for her statistical expertise and support.

Most importantly, I would like to thank my advisor and mentor, Dr. Edward Shapiro, who has provided me with encouragement, support, and feedback throughout this project and my entire graduate school career. Thank you for affording me the opportunity to work on several of your research projects which inspired my dissertation

topic. Thank you for always challenging me to grow as a professional and helping me hone my skills as a scientist-practitioner.

I would like to thank my friends and family for believing in me and supporting me throughout my graduate training. Thank you to my Lehigh support system of Laura Rutherford, Jilda Hodges Ulicny, Kimberly Seymour, Erin Leichman, and Brigid Vilardo; I am eternally grateful for your unwavering support and encouragement. Thank you to my friends Alexis Cohen, Lauren Kaplan, and Melissa Kaufman for always making me smile, providing a diversion from work, and understanding when work prevented me from joining in the festivities. I would also like to thank my nephew, Hudson, for brining immense joy to my life, and my extended family (Scott, Rachael, Andy, Erin, Brad, Mary, and Richard) for their support and patience with my graduate studies. To my beloved late mother (Barbara) and grandmother (Marge) thank you for your unconditional love and support. Your memory has served as a reminder of the endurance of the human spirit in spite of life's obstacles. To my father, thank you for instilling in me a passion for education and determination to reach my goals. Your unending love and support has helped me to reach this milestone in my career.

Last but not least, I am eternally grateful to my husband, Brian Thomas, without whom I would not be here today. Thank you for providing 24/7 technical and emotional support. Your love, patience, understanding, and sense of humor helped me to navigate the rocky terrain and reach the pinnacle of my graduate studies. For these reasons, I dedicate my dissertation to my beloved husband, Brian Thomas.

Table of Contents

List of Tables

List of Figures

Abstract

Reading comprehension is a critical aspect of the reading process. Children who experience significant problems in reading comprehension are at risk for long-term academic and social problems. High-quality measures are needed for early, efficient, and effective identification of children in need of remediation in reading comprehension. Substantial effort has been devoted to developing measures for identifying children at risk for reading difficulties; however, the optimal combination of measures has not been determined. One method that has been considered as having potential for assessing reading comprehension is retelling. The purpose of this study was to examine the technical adequacy and usability of an oral retelling procedure that employed a rubric scoring method to assess the reading comprehension of students in third grade. This study investigated the convergent and predictive validity of the Reading Retell Rubric (RRR) for identifying children at risk for reading comprehension difficulties on summative reading assessments. Reading data from curriculum-based measures of oral reading and comprehension of narrative and expository text and criterion-measures of reading comprehension and overall reading ability were gathered from 107 elementary school children attending third grade in a public elementary school. Results indicated that participants demonstrated greater comprehension for narrative text. This investigation reinforces the strength of ORF in predicting reading ability on summative assessments. More research is needed to determine the usability of the RRR. Findings suggest that the RRR may be a viable alternative to the Adapted Retell Fluency measure. In addition, it is speculated that the RRR may be useful as a diagnostic tool (instead of a universal screener) within a multiple-gated screening process.

Chapter One: Statement of the Problem

Roughly 10 million children (~17.5%) in the United States will experience problems related to reading within their first three years of schooling (National Reading Panel [NRP], 2000). Research has shown that students who fail to learn how to read by the end of third grade will continue to have significant impairments well beyond this period if they do not receive appropriate intervention (Cain & Oakhill, 2006b; Foorman, Francis, Shaywitz, Shaywitz, & Fletcher, 1997; Francis, 1996; Juel, 1988; Simmons, Kuykendall, King, Cornachione, & Kame'enui, 2000). Without intervention, reading deficits become more pronounced as students advance through the grades and are required to read more difficult material, such as content area, expository text (Taylor, Alber, & Walker, 2002). As students fall further behind their peers, it becomes increasingly less likely that they will catch up (Bursuck & Damer, 2007; Stanovich, 1986).

Reading deficits have a significant impact not only for the individual but also society; indeed, children who experience significant difficulties in reading are at risk for long-term academic and social problems including truancy, high school dropout, teen pregnancy, substance abuse, delinquency, and incarceration (Burke & Hagan-Burke, 2007; McGill-Franzen, 1987). According to Vanderstaay (2006), "the likelihood that a child will commit a delinquent act rises as school performance declines and falls as school performance improves" (p. 331). Of the children who experience significant problems in reading, nearly 10 to 15 percent eventually drop out of high school and only 2 percent complete a four-year college program (Whitehurst & Massetti, 2004).

Prevention or remediation of difficulties associated with reading deficits requires valid, reliable, and sensitive tools for early identification of children at risk for reading problems, in particular students who are likely to perform poorly on statewide reading assessments. With the inception of the No Child Left Behind Legislation (NCLB; 2002, PL 107-110), schools are held more accountable for students' progress. Outcomes from the statewide reading assessment have important implications for state and district-level decision making and policy, which in turn, have significant ramifications for individual students, teachers, schools, and districts (Shapiro, Solari, & Petscher, 2008). The increased accountability created by NCLB has motivated a paradigm shift from a "wait-to-fail" model to an early identification model for detecting students in need of intervention. Alternative to the "wait-to-fail" model, early identification through systematic screening provides opportunities for remediation prior to failure; thereby, preventing the development of more severe reading difficulties and reducing the incidence of academic and/or behavioral problems. In particular, a vast body of research has indicated that through early identification and intervention many students who experience early reading problems can become competent readers (Denton, Fletcher, Anthony, & Francis, 2006; Foorman et al., 1997; Scanlon, Vellutino, Small, Fanuele, & Sweeney, 2005).

There are five key skills that constitute the construct of reading. These include phonological awareness, alphabetic principle, fluency, vocabulary, and comprehension (NRP, 2000). Reading develops in a series of distinct stages from prereading (phonological awareness) to learning-to-read (alphabetic principle and fluency) to reading-to-learn (vocabulary and comprehension) (Chall, 1983). The ultimate goal of reading is comprehension, which involves the ability to derive meaning from written text. Struggling readers often have difficulty with aspects of reading comprehension including attending to the meaning of the text, remembering facts, identifying the main ideas, drawing on prior knowledge, making inferences, and monitoring their comprehension (Taylor, Alber, & Walker, 2002). Comprehension is a crucial skill that should be included in an assessment for identifying children at risk for reading difficulties (RAND Reading Study Group [RRSG], 2002; Sweet, 2005).

Substantial effort has been devoted to developing measures for identifying children at risk for reading comprehension difficulties; however, "a satisfactory solution has yet to emerge" (Speece, 2005, p. 487). In part, this is likely due to the multifaceted nature of reading, which makes it challenging to assess. Standardized, norm-referenced tests are often used to assess overall reading skills and provide an indicator of a student's performance compared with same-age and grade-level peers. However, standardized, norm-referenced tests have been criticized for (a) indirectly measuring academic skills contained within a curriculum, (b) ignoring the importance of fluency, (c) failing to provide instructionally useful information, and (d) being problematic for progress monitoring of student growth over time (e.g., Barnett, Lentz, & Macmann, 2000; Elliott, Huai, & Roach, 2007; Gresham & Witt, 1997; Klingner, 2004; Marston, 1989). For these reasons, standardized, norm-referenced reading tests may lack usefulness as a screening measure.

Alternatively, curriculum-based measurement (CBM) was designed for measuring students' academic proficiency in basic skill areas to determine when instructional modifications are necessary in general or special education and to monitor an individual's response to intervention (e.g., Christ & Silberglitt, 2007; Deno, 1985, 1986, 1989; Deno & Mirkin, 1977; Fuchs & Deno, 1991; Fuchs & Fuchs, 2002; Shinn, 1989, 1998, 2002). Compared to standardized, norm-referenced tests, CBM demonstrates utility as a screening method because it is cost and time efficient, directly assesses specific skills that are indicators of overall performance in a basic skill area, allows for repeated measurement over short periods of time, and demonstrates sensitivity to short-term changes in performance (Watson & Skinner, 2004).

Research on the application of CBM for early identification of reading difficulties has predominately examined the assessment of Oral Reading Fluency (ORF). ORF directly assesses the speed and accuracy of oral production of text (Adams, 1990; Berninger, Abbott, Billingsley, & Nagy, 2001). Students typically read a short narrative or expository passage from grade-level controlled reading material. The number of words read correctly per one minute (WCPM) comprises the student's performance score. Although on the surface it appears to be measuring only speed and accuracy of oral reading, research has provided strong support for ORF as an indicator of overall reading proficiency, including comprehension. For example, Fuchs, Fuchs, and Maxwell (1988) found ORF scores to correlate higher with a criterion measure of reading comprehension (r = .92) than decoding (r = .81). A consistent pattern has emerged in the research on ORF used with students in first through sixth grades in which ORF scores have demonstrated a moderate to strong relationship with criterion measures of reading comprehension with most correlations around .65 (Reschly, Busch, Betts, Deno, & Long, 2009).

Research has also found ORF scores in the early grades to be reasonable predictors of comprehension (Rasinski, 1990), and an effective tool for identifying students in need of additional reading instruction (Fuchs & Fuchs, 1992; Jenkins & Jewell, 1993). In particular, research has found ORF scores to be a significant predictor of students' performance on statewide reading assessments. For example, on the Dynamic Indicators of Basic Early Literacy Skills (DIBELS, 6th edition; Good & Kaminski, 2002), at third grade, 110 WCPM and higher has been identified as an appropriate cut score for predicting success (i.e., low risk) on statewide reading assessments. Ninety-one percent of students who achieved this cut score were found to be successful on the reading portion of the Florida Comprehensive Assessment Test – Sunshine State Standards (Buck & Torgeson, 2003).

However, research has found that ORF alone is not a strong indicator of reading comprehension as students advance in grade level; by fourth and fifth grade, ORF may not offer as much sensitivity for measuring students' reading comprehension skills as other measures (Hosp & Fuchs, 2005; Kranzler, Miller, & Jordan, 1999; Shapiro et al., 2008; Shinn et al., 1992). For example, Jenkins and Jewell (1993) found the relationship between ORF and criterion measures of overall reading ability and comprehension to decrease from Grades 2 to 6 (e.g., Gates-MacGinitie, Total Reading: Grade 2 r = .83 to Grade 6 r = .67; Comprehension: Grade 2 r = .86 to Grade 6 r = .63). Contrary to early elementary in which ORF alone has been identified as a strong indicator of overall reading ability, Speece et al. (2010) found that the overall reading ability of the fourth grade students in their sample could best be explained by a multivariate model which included comprehension, word reading, and fluency. Given that the predictive power of

ORF appears to decrease as grade level increases, direct measurement of comprehension is needed, especially for students in higher grade levels. Adding a measure of reading comprehension to ORF will likely enhance the decision making process for identifying proficient and non-proficient readers.

An additional reason for including a direct measure of reading comprehension is to increase the face and content validity of the assessment. Despite strong empirical support for ORF as an indicator of students' overall reading proficiency, a primary barrier to acceptance among teachers of ORF is its lack of face validity as a measure of reading comprehension (Roberts, Good, & Corcoran, 2005; Shinn et al., 1992; Williams, Skinner, Floyd, Hale, Neddenriep, & Kirk, 2011). ORF requires students to read text aloud and does not directly assess students' understanding of what they read; consequently, teachers may not view ORF as having the needed face validity to accept it as a measure of reading comprehension or content validity to design interventions for students who struggle with comprehension. In particular, practitioners report concern with ORF's ability to detect "word callers" (Dewitz & Dewitz, 2003; Hamilton & Shinn, 2003; Meisinger, Bradley, Schwanenflugel, Kuhn, & Morris, 2009; Roberts et al.; Shapiro, 2004), which are readers who have fluent decoding without high levels of comprehension (Stanovich, 1986). Adding a measure of reading comprehension to ORF will likely enhance the decision making process for detecting "word callers" and inform intervention development by assisting in identification of comprehension skill deficits (e.g., sequencing).

One potential method for enhancing ORF's measurement of reading comprehension is Free Oral Retell. After reading an entire passage, the passage is removed from view, and the student is asked to retell the key information from the passage in his or her own words (e.g., what the passage was all about). The term “free” indicates that the oral retell is not prompted, meaning that no cues are provided to assist the individual in retelling the text. Free oral retell measures a broad range of comprehension skills that are directly linked to instruction and intervention (Klinger, 2004; Roberts et al., 2005). In particular, free oral retell provides a view of the quantity, quality, and organization of information a reader ascertained from reading the text (Winograd, Wixson, & Lipson, 1989), including a student’s understanding of the passage, memory of events, and ability to sequence events and major concepts (Hansen, 1978; Ringler & Weber, 1984). Free oral retell shows potential as a screening method for identifying students in need of intervention because it is time efficient to create and administer and it yields a large sample of comprehension behaviors that can inform intervention and increase the chance of detecting post-intervention changes (Roberts et al.).

One barrier to the use of free oral retell as a screening measure is efficiency of scoring. Although retelling has frequently been used as an assessment tool of reading comprehension in reading research (Fuchs et al., 1988; Gambrell, Pfeiffer, & Wilson, 1985; Johnson, 1983; Schisler, Joseph, Konrad, & Alber-Morgan, 2010), it has less frequently been used for assessment in applied settings (Blachowicz & Ogle, 2008; Maria, 1990). This is likely due to the scoring of retells, which can be cumbersome and time-consuming (Fuchs & Fuchs, 1992; Johnston, 1982; Klingner, Vaughn, & Boardman, 2007). Typically, researchers use a text analysis system to divide the passage into idea units (i.e., propositions) and assign idea units a particular level of importance. The oral retell is transcribed with two independent scorers determining the number of idea units identified in the oral retell. This method is not realistic or feasible for school settings; it is unlikely that a teacher would have enough training in using the text analysis scoring system or time to have two independent teachers transcribe and score each student's oral retell (Maria).

In order to identify the most valid and feasible method for scoring oral retells, Fuchs and colleagues (1988) examined the following scoring methods: (a) counting the total number of words retold (b) calculating the percentage of content words retold, and (c) calculating the percentage of idea units retold. In this study, participants were allotted 5 minutes to read a 400-word folktale passage and 10 minutes to retell the passage. Participants' retells were audio recorded and transcribed for scoring. Results from their study revealed that the different methods of scoring oral retell related comparably with each other (*r* = .84 to .94) and had similar correlations with the other measures of reading ability, including the Reading Comprehension subtest of the Stanford Achievement Test (SAT-7; Gardner et al., 1982, 1983) (*r* = .59 to .64), with the most feasible scoring method being counting the total number of words retold (Fuchs et al.). Despite the feasibility over the other scoring methods, counting the total number of words retold required transcription of the student's retell and multiple scorers, thereby, decreasing the feasibility of this method for the classroom setting.

Alternatively, Good and Kaminski (2002) as part of the DIBELS 6th edition and Roberts et al. (2005) as part of the Vital Indicators of Progress (VIP) used within the Voyager Universal Literacy System have developed a method for scoring a retell that does not require use of a text analysis scoring system or transcription. Participants are allotted 1 minute to read aloud a 200 or more word grade-level controlled passage. After this 1 minute time period, the passage is removed, and the participant is given 1 minute to retell what they just read in their own words. As the participant is orally responding, the examiner records the total number of words that the participant can retell within the one minute time period. The retell fluency (RTF) measure was specifically designed to complement ORF, as a means of improving ORF's face validity and accuracy of detecting "word callers" (Good & Kaminski).

As previously mentioned, free oral retell can provide a view of the quantity, quality, and organization of information that a reader amassed from reading a text (Maria, 1990; Winograd et al., 1989). However, the RTF measure in the DIBELS 6th edition is limited in that it only records information regarding the quantity of the retell. In particular, mistakes or inconsistency in the retell do not count against the student as long as the student is fundamentally on topic (Good & Kaminski, 2002). Consequently, a lengthy retell with several inaccuracies could score a false negative, mistakenly placing the individual in the low risk range due to the high number of words retold whereas a short and concise retell that includes the key story structure elements could score a false positive. In addition, there are no scoring guidelines for the exact number of words that should be included in an effective retell when using the DIBELS 6th edition RTF measure, thus limiting interpretations of the RTF score, progress monitoring, and instructional decision making.

Research investigating the psychometric properties of the DIBELS 6th edition RTF measure has found ORF to be a better predictor of comprehension and more strongly correlated with measures of reading skills and comprehension than RTF (Dynamic Measurement Group [DMG], 20110b; Marcotte & Hintze, 2009; McKenna &

Good, 2003; Pressley, Hilden, & Shankland, 2005; Riedel, 2007). Furthermore, Marcotte and Hintze found the DIBELS 6th edition RTF measure to consistently yield the lowest correlations (range, *r* = .45–.49) with other measures of reading comprehension (i.e., GRADE, ORF, sentence verification technique, maze, and written retell) and insignificantly contribute to the prediction of reading comprehension (i.e., GRADE) for a sample of fourth grade students. In an examination of the DIBELS 6th edition RTF measure with a group of first and second graders, Riedel concluded that there is a "lack of empirical evidence for the usefulness of the RTF task" (p. 560). In this study, RTF was found to be a weaker predictor of comprehension than ORF and did not substantially improve the predictive accuracy compared to ORF alone (Riedel). Consequently, the DIBELS 6th edition RTF measure lacks utility as a screening measure and may not be meeting its intended purpose of serving as an indicator of comprehension.

Perhaps a better method for judging an effective retell may include evaluating the accuracy of components, sequence, and coherence of the retell in which students are awarded points for each idea or fact recalled. Scoring based on the quality and organization of information amassed from the passage allows for greater conceptual match to what we know are the key elements of reading comprehension, including a reader's understanding of the story structure and ability to sequence information. There is limited research on the use of story structure elements as the methodology for scoring a participant's retell. Several different scoring methodologies were identified in the literature including: (a) counting the total number or proportion of story elements included in the retell (Gambrell, Koskinen, & Kapinus, 1991), (b) awarding varying point amounts for including specific story elements in the retell (Hagtvet, 2003), (c) awarding points on a sliding scale based on the amount of information provided for the specific story element in the retell (Shannon, Kame'enui, & Baumann, 1988; Short, Yates, & Feagans, 1992), and (d) scoring the quality of information provided for the specific story element in the retell (Rabren, Darch, & Eaves, 1999). However, none of these studies have investigated the psychometric properties of using story structure elements to score free oral retell.

The creators of the DIBELS recently released a revised version of the RTF measure within the newly published DIBELS Next (DMG, 2011a) called DIBELS Oral Reading Fluency Retell (DORF Retell). The DORF Retell continues to provide information regarding the quantity of the retell (i.e., the total number of words *related to the passage* that the participant can retell within the one minute time period); however, the DORF Retell is an improvement on the DIBELS 6th edition RTF measure because it also provides information regarding the quality and sequence of information recalled. The quality and sequence of information recalled are assessed through the newly added (a) "Quality of Response" rating, which requires the examiner to indicate from 1 to 4 how many details related to the main idea were provided in the retell and whether the details were provided in a meaningful sequence (i.e., 1 = *provides 2 or fewer details*; 2 = *provides 3 or more details*; 3 = *provides 3 or more details in a meaningful sequence*; or 4 = *provides 3 or more details in a meaningful sequence that captures a main idea*) and (b) "General Retell Response Patterns" checklist, which requires the examiner to record whether across all three passages the participant: *summarizes, repeats the same detail, retells the passage verbatim, "speed reads" the passage and has limited retell relative to number of words read,* or *talks about own life related to passage* (DMG, 2011a). In addition, the DORF Retell now includes benchmark goals and cut points for (a) the exact number of words *related to the passage* that should be included in an effective retell and (b) the quality of response rating that reflects a student's qualitative understanding of the passage, thus allowing for interpretation of the DORF Retell score. Consistent with research on the technical adequacy of the DIBELS 6th edition RTF measure, in the technical manual for the DIBELS Next, the creators of the DORF Retell measure report more consistent and higher correlation coefficients for reliability and validity of the DORF WCPM compared to DORF Retell score (DMG, 2011b; Powell, Smith, Good, Latimer, Dewey, & Kaminski, 2011). Note the DORF Retell was released after this dissertation research study was conducted.

Reed (2011) examined the psychometric properties of the DORF Retell along with 10 other commercially or publically available retell measures. All but one measure (i.e., VIP) examined the story ideas or facts recalled. Reed concluded that all of the retell measures reviewed provided insufficient information regarding the psychometric properties of the instruments resulting in a lack of confidence in the existing retell measures' ability to assess students' reading comprehension and inform intervention development. In particular, Reed indicated that future research should seek to improve both the technical adequacy and practical relevance of retell measurement in order to possess instructional utility.

Given the limited and varied research on using story structure elements as the methodology for scoring a participant's retell, future research was warranted to examine the (a) contextual appropriateness (e.g., alignment with constructs of interest and fit with population of interest), (b) technical adequacy (e.g., reliability and validity), and (c) usability (e.g., cost, feasibility, efficiency, acceptability, and utility of outcomes) (Glover & Albers, 2007) of utilizing a rubric method for scoring free oral retell. Two preliminary studies have examined the use of the Reading Retell Rubric (RRR) for measuring reading comprehension of key narrative text story structure elements (Shapiro, Fritschmann, Thomas, Hughes, & McDougal, 2010) and key expository text story structure elements (Fritschmann, Shapiro, & Thomas, 2010) compared to an adapted version of the DIBELS 6th edition RTF measure. The adapted RTF measure combined elements of the Fuchs et al. (1988) method for calculating the total number of words retold and the Good and Kaminski (2002) and Roberts et al. (2005) method for scoring RTF. Specifically, participants were permitted to finish reading the entire passage before retelling the story for 1 minute. In addition, the passage remained in view during the retell.

The RRR for narrative text and expository text were developed through a review of the literature and examination of story elements that could be identified in commercially available oral reading probes. For example, according to Caldwell and Leslie (2005) a good narrative retell includes the major story elements (e.g., characters, goal/problem, events, resolution), is sequential, and makes causal connections between events in the story whereas a good expository retell is guided by knowledge of the topic and expository text structure, is retold in a sequential or time-ordered format, identifies important information (e.g., main idea and details), and may include cause and effect, problem and solution, or compare and contrast. Narrative and expository texts differ in person (e.g., narrative texts are generally about people or characters and written from a personal perspective), orientation (e.g., expository texts are subject-oriented), time (e.g., narrative texts link events in a chronological order), and linkages (e.g., expository texts link events in a logical order) (Copmann & Griffith, 1994); consequently, the RRR for narrative text consists of different items than the RRR for expository text. Specifically, for the narrative version of the RRR students could earn a total of 10 points for correctly providing each of the following story elements in their retell: theme, problem, goal, setting, characteristic, initiating event, climax, sequence, problem solution, and ending of the story, whereas for the expository version of the RRR students could earn a total of 10 points for correctly providing each of the following content in their retell: topic, main idea, primary supporting details (up to 4 points), and secondary support details (up to 4 points).

The initial investigation of the Narrative version of the RRR was conducted with a different sample than the investigation of the Expository version of the RRR, thus limiting direct comparisons of text type. Results of the convergent validity analyses across the third grade narrative and expository studies were mixed; differences were noted in the magnitude and significance of correlations between RRR with ORF and Pennsylvania System of Student Assessment (PSSA; Data Recognition Corporation [DRC], 2009). Across both studies the RRR had the highest correlations with Adapted RTF (Narrative winter r = .59 & spring r = .42, p < .01; Expository winter r = .55 & spring r = .46, p < .01) and weaker correlations with ORF (Narrative winter r = .23 & spring r = .21, p < .01; Expository winter r = .16 & spring r = .12, ns) and PSSA (Narrative winter r = .24 & spring r = .25, p < .01; Expository winter r = .02 & spring r = -.02, ns). Differences were also noted in the backwards elimination regression analysis for variables predicting third grade PSSA scores. For the narrative study RRR added significantly (p < .05) to ORF's (p < .001) prediction of PSSA, with ORF and RRR accounting for 30% of the variance in explaining PSSA. Conversely, for the expository study, RRR did not add significantly to ORF's (*p* < .001) prediction of PSSA, with ORF alone accounting for 31% of the variance in explaining PSSA. Note, it was speculated that having the passage present during the retell may have impacted the findings, with some participants more likely to copy directly from the text as opposed to engaging in more active and deeper processing (Hidi & Anderson, 1986).

Further research is warranted to examine the psychometric and diagnostic properties of the RRR measure to determine its usefulness as a screening measure of reading comprehension abilities. The reasons for this are (a) the need for valid, reliable, and sensitive measures for identifying children at risk for reading problems, (b) the limitations of standardized, norm-referenced reading tests for screening and instructional decision making, (c) the limitations of ORF in assessing reading comprehension, (d) the limitations of existing retell measures including weak technical adequacy and instructional utility, and (e) the preliminary nature and mixed findings across the two previous investigations of the RRR. The purpose of this investigation was to evaluate the utility of using the RRR for identifying children at risk for reading comprehension difficulties with narrative and expository text. In particular, this study sought to replicate and expand on the previous investigations of the RRR. This investigation also examined the convergent validity of the RRR by comparing performance on the RRR with performance on other established measures of reading comprehension administered at the same point in time. However, this study broadened the scope of the Shapiro et al. (2010) and Fritschman et al. (2010) studies by allowing for direct comparisons of text type within the same study and examining the classification accuracy of the RRR. Also, unlike in the previous investigations, the passage was removed from view during the retell.

Research Questions and Hypotheses

RQ1: What is the convergent validity of the RRR for assessing reading comprehension of narrative text with other measures typically used to assess narrative reading comprehension (ORF, Adapted RTF, Group Reading Assessment and Diagnostic Evaluation [GRADE; Williams, 2001], 4Sight Reading Benchmark Assessment [4Sight; Success for All Foundation, 2008], and PSSA)?

RQ2: What is the convergent validity of the RRR for assessing reading comprehension of expository text with other measures typically used to assess expository reading comprehension (ORF, Adapted RTF, GRADE, 4Sight, and PSSA)?

> H1 & H2: Similar to the Fuchs et al. (1988) study in which different methods of scoring oral retell were highly correlated, it was hypothesized that the RRR for narrative and expository texts would be highly correlated with the Adapted RTF. This notion was also supported by the findings from the preliminary investigations of the RRR, which yielded the highest correlations between RRR and RTF. It was also hypothesized that the RRR for both narrative and expository texts would have low to moderate correlations with ORF, GRADE, 4Sight, and PSSA. This hypothesis was supported by findings of the DMG (2008, 2011b), who reported moderate correlations between RTF and ORF and between RTF and GRADE, and Fuchs et al. who found the methods of scoring oral retell to yield moderate correlations with the total number of words read correctly and with a standardized, norm-referenced measure of reading comprehension. Research on

the assessment of narrative and expository text has yielded different results across measures, with students consistently achieving higher scores on narrative text (Pearson & Hamm, 2005). Therefore, it was hypothesized that the correlations between measures of narrative text would be higher than those between measures of expository text.

RQ3: Does the RRR for assessing reading comprehension of narrative text improve ORF's prediction of students who are proficient readers and those who have been identified as non-proficient readers on the (a) GRADE, (b) 4Sight, and (c) PSSA?

RQ4: Does the RRR for assessing reading comprehension of expository text improve ORF's prediction of students who are proficient readers and those who have been identified as non-proficient readers on the (a) GRADE, (b) 4Sight, and (c) PSSA?

H3 & H4: It was hypothesized that the combination of the RRR for narrative text and ORF would be the strongest prediction model, with the RRR significantly adding to the accurate prediction of non-proficient and proficient readers on the 4Sight, PSSA, and GRADE. Despite findings from the preliminary investigation of the RRR expository, which found RRR expository did not add significantly to ORF's prediction of the PSSA, it was hypothesized that the combination of the RRR for expository text and ORF would be a stronger prediction model than ORF alone, with the RRR significantly adding to the accurate prediction of non-proficient and proficient readers on the 4Sight, PSSA, and GRADE. The preliminary study utilized a backwards elimination regression technique in which ORF, Adapted RTF, and RRR were included in the same analysis. Several problems have been identified with backwards elimination regression. In backwards elimination regression the order of elimination is based solely on the empirical relationship among the variables entered into the equation. As noted by Licht (1995) “pure empirical selection of predictors is likely to be highly sample specific and is not likely to include all theoretically relevant, or to exclude all irrelevant predictors. Thus, these procedures are likely to produce misleading and nonreproducible results” (p.53). Alternatively, this study used Hierarchical Binary Logistic Regression analysis to examine the predictive validity of the ORF and RRR measures. Logistic Regression has been widely used in the medical literature and has increased in use in the social science and educational research (Peng, Lee, & Ingersoll, 2002). Using Hierarchical Binary Logistic Regression allowed for examination of ORF’s ability, with the additive benefit of the RRR, to predict reading performance (proficient or non-proficient) on the dependent variables.

RQ5: Did the RRR for assessing reading comprehension of narrative text have a greater contribution to ORF’s prediction of students who are proficient readers and those who have been identified as non-proficient readers on the (a) GRADE, (b) 4Sight, and (c) PSSA, as compared to Adapted RTF?

RQ6: Did the RRR for assessing reading comprehension of expository text have a greater contribution to ORF’s prediction of students who are proficient readers and those who have been identified as non-proficient readers on the (a) GRADE, (b) 4Sight, and (c) PSSA, as compared to Adapted RTF?

H5 & H6: It was hypothesized that the combination of the RRR (narrative or expository) and ORF would be a stronger prediction model than the combination of Adapted RTF and ORF. This was based on the notion that the RRR measure yields a larger sample of comprehension behaviors as compared to the RTF measure, which only provides information regarding the quantity of an individual’s retell.

Chapter Two: Literature Review

After conducting an exhaustive search, Bishop (2003) concluded that the ability to predict the children most at risk for reading comprehension problems has not been perfected and the optimal combination of measures has not been determined. This is likely due to the complex and multifaceted nature of reading comprehension, which makes it challenging to assess (RAND Reading Study Group [RRSG], 2002). Despite these challenges, there is a strong need to develop valid, reliable, and sensitive screening measures to identify children at risk for reading problems, in particular students who are likely to perform poorly on statewide reading assessments.

This chapter begins with a brief overview of reading development in order to provide a foundation for understanding reading comprehension. The characteristics of reading comprehension are reviewed to provide a framework for understanding the complex and multifaceted nature of reading comprehension. Next, an overview of the rationale for early identification of students who struggle with reading comprehension is provided, followed by a discussion on the measurement of reading comprehension. Limitations of current methods for assessing reading comprehension are offered. Finally, the Reading Retell Rubric (RRR) is presented as an alternative to the existing methods of assessing reading comprehension.

Brief Overview of Reading Development

Chall (1983) conceptualized reading as a series of distinct stages of development that the reader progresses through as he or she becomes a more proficient reader. Each of the six stages are qualitatively different, spanning skills from prereading to learning-to-read to reading-to-learn. The prereading stage, birth to age 6, includes the development of oral language, visual and visual-motor skills, and auditory perceptual skills in which the child begins to gain control over language, insight into print, and letter recognition (Chall). Grades 1 and 2 span the initial reading or decoding stage, which includes acquisition of the alphabetic principle, sound-spelling relationships, decoding skills, and recognition of printed words (Chall). Confirmation, fluency, and ungluing from print mark the skill development at Grades 2 and 3, which includes using decoding skills to confirm what is already known, decoding unknown words, increasing reading speed and accuracy, and gaining insight into comprehension of text (Chall).

A shift in reading occurs at Grades 4 through 8, in which students begin to read in order to learn new information (Chall, 1983). During this stage, reading is viewed as a tool for acquiring knowledge, which includes locating information in text, expanding vocabularies, and building on prior background and world knowledge (Chall). High school reading requires the understanding of multiple viewpoints, development of more complex language and cognitive abilities, and ability to critically analyze text (Chall). Finally, at the college level and beyond, readers construct and reconstruct their understanding of text based on analysis and synthesis skills and are able to select printed material for the purpose of constructing knowledge (Chall).

As analogized by Margolis (2004) “reading is like a car’s engine [in that] all parts must work simultaneously and smoothly in logical coordination with one another for the car to go” (p. 195). If a reader fails to master skills from the previous stage, a “snowballing effect” will likely occur, leading to significant reading problems and an inability to keep-up with grade-level expectations (Stanovich, 1986). Consequently, reading requires the coordination and integration of phonological awareness, alphabetic principle, fluency, vocabulary, and comprehension (National Reading Panel [NRP], 2000). For example, if a student has difficulty noticing, thinking about, and working with individual sounds in words (i.e., phonological awareness) then he or she will likely have problems with the alphabetic principle (e.g., acquisition of letter-sound correspondence), which will impact the ability to use decoding skills (e.g., a word analysis skill readers use to pronounce a word when it is not recognized instantly; Ekwall & Shanker, 1989) and therefore impede fluency (e.g., the ability to read quickly, accurately, and with appropriate expression; NRP) and understanding of vocabulary (e.g., the ability to understand and use words to acquire and convey meaning; NRP) and thus reading comprehension (Margolis).

Factors Affecting Reading Comprehension

The ultimate goal of reading is comprehension. Although the elements of phonological awareness, alphabetic principle, decoding, fluency, and vocabulary are essential to the reading process, “if there is no comprehension, there is no reading” (Durkin, 1980, p. 191). Reading comprehension involves the ability to derive meaning from written text through answering or generating questions, demonstrating understanding of story structure, monitoring comprehension, retelling and summarizing, and analyzing text through making predictions, connections, and inferences (University of Oregon Center on Teaching and Learning, 2010). The NRP (2000) defines reading comprehension as a complex cognitive process that draws on vocabulary skills and requires an intentional and thoughtful interaction with the text. Reading comprehension draws on both lower-level lexical skills (e.g., word reading efficiency, vocabulary knowledge, and knowledge of grammatical structure) and higher-level text processing skills (e.g., inference generation, comprehension monitoring, and working memory capacity) (Cain & Oakhill, 2006a). Efficiency in processing of lower-level lexical skills can facilitate reading comprehension by freeing-up more resources for higher-level text processing skills (Cain & Oakhill). As a result, reading comprehension requires both bottom-up (e.g., identification of printed words) and top-down (e.g., understanding of semantic and syntactic relationships among words) processing in order to accurately comprehend text (Cutting & Scarborough, 2006).

According to the RRSG, reading comprehension is the “process of simultaneously extracting and constructing meaning” (Snow & Sweet, 2003, p. 1), which revolves around the interaction between the reader, text, and environment. When comprehension does not proceed smoothly, it is likely due to a break down in the interface between the knowledge that a reader brings to the text, the reader’s interpretation of the text, and the situation in which the text is read.

Reader factors. A reader’s phonological awareness, decoding ability, reading fluency, vocabulary knowledge, world knowledge, attention, memory, interest, motivation, self-efficacy, and analysis, inference, visualization, and metacognitive skills can all impact comprehension (Maria, 1990; Rapp, van den Broek, McMaster, Panayiota, & Espin, 2007; Snow & Sweet, 2003). As defined by Gough’s “simple view of reading” (Gough, 1996; Hoover & Gough, 1990), reading comprehension is the product of recognizing words on the page and understanding the words once they have been recognized. For this reason, a reader’s decoding abilities and vocabulary knowledge are crucial to the comprehension process. In particular, weakness in decoding abilities is the most common and debilitating source of reading difficulties (Adams, Foorman, Lundberg, & Beeler, 1998; Juel, 1991; Nation, 2005; Shankweiler et al., 1995; Stanovich, 1986; Vellutino, 2003). Readers need to "conquer the code in order to master the meaning" (Cohen, 1996, p. 76). Slow or inaccurate decoding skills can interfere with comprehension by inhibiting a reader's connection with the text and memory for events, and consequently, poor word identification skills are strongly correlated with poor reading comprehension skills (Adams et al.; Rack, Snowling, & Olson, 1992; Stanovich, 1992; Vellutino). A reader's vocabulary knowledge (e.g., semantic and syntactic) is also a strong predictor of reading comprehension (Oakhill & Cain, 2007). As the breadth (e.g., the number of words with known meaning) of a reader's vocabulary knowledge increases, so does the depth (e.g., the richness of knowledge about words that are known), thus permitting flexibility in his or her understanding and use of word meanings (Tannenbaum, Torgesen, & Wagner, 2006) and use of sentence structure to supplement decoding ability (Tunmer & Hoover, 1992).

Interaction between reader and text factors. Difficulties with reading comprehension are particularly exacerbated when there is a poor match between a reader's knowledge and text factors (e.g., topic, source, and readability level) (Rapp et al., 2007; Snow & Sweet, 2003). Research has consistently shown that readers with more world knowledge and/or interest in the *topic* have a better understanding and retention of story elements (Medina & Pilonieta, 2006; Pressley, 2000; Snow & Sweet, 2003). Readers with more knowledge of a topic area are more likely to be interested in the passage, and thus more motivated to read the passage (Kintsch & Kintsch, 2005). Therefore, knowledge of the topic can facilitate a reader's interest, motivation, attention, memory, understanding, and ability to make inferences about the text (Caldwell, 2008;

Snow & Sweet, 2003). This is particularly salient for expository text, which tends to have greater demands for background knowledge. For example, Best, Floyd, and McNamara (2008) investigated the effects of text genre, decoding skills, and world knowledge on third graders' text comprehension. Results revealed that comprehension was better for narrative text than expository text, with expository text comprehension greatly influenced by world knowledge (Best et al.).

Consequently, the passage *source*, whether the passage was drawn from narrative or expository text can also impact text comprehension. Whereas narrative texts tend to follow a predictable structure or sequence of events, expository texts tend to have greater structural complexity (Best et al., 2008). In particular, narrative and expository texts differ in person (e.g., narrative texts are generally about people or characters and written from a personal perspective), orientation (e.g., expository texts are subject-oriented), time (e.g., narrative texts link events in a chronological order), and linkages (e.g., expository texts link events in a logical order) (Copmann & Griffith, 1994). Research has consistently yielded greater comprehension for narrative text (Pearson & Hamm, 2005), which may be attributed to the prior knowledge and vocabulary demands of expository text, which is not as predictable, consistent, or as clear. Specifically, research has found that a reader's understanding of expository text structure can aid in comprehension (Hall, Sabey, & McClellan, 2005); as a result, readers who lack knowledge about expository text structure will likely have difficulty organizing and processing text content (Best et al.).

Elementary-school children often encounter greater difficulty comprehending expository text as compared to narrative text (Duke & Kays, 1998; Olson, 1985; Spiro &

Taylor, 1980). For example, in Greece, Diakidoy, Stylianou, Karefillidou, and Papageorgiou (2005) found for both listening and reading comprehension, expository comprehension levels were lower than narrative comprehension levels across grades 2 to 8, with expository text comprehension steadily increasing from grades 2 to 8. Students with reading deficits have also been found to have more difficulty comprehending expository text (e.g., Gajria, Jitendra, Sood, & Sacks, 2007; Gersten, Fuchs, Williams, & Baker, 2001; Johnson, Graham, & Harris, 1997; Warren & Fitzgerald, 1997; Williams, 2005). For instance, in an examination of the reading-related science skills of fourth and sixth grade students with and without learning disabilities, Carlisle (1993) found that students with learning disabilities performed significantly weaker on expository text comprehension as compared to non-learning disabled counterparts.

Difficulties with reading comprehension are also exacerbated when there is a mismatch between the reader's abilities and text features, in particular when a reader is asked to read material that is above his or her instructional *reading level* (Compton, Appleton, & Hosp, 2004; Stanovich, 1986). The text can be difficult or easy depending on the match between the text factors and a reader's knowledge, experience, and abilities (Snow & Sweet, 2003). Text readability includes the vocabulary, familiarity of words, sentence length, and coherence of the passage (Francis et al., 2008). In particular, passages with too many unfamiliar words, complex syntax patterns, and a high level of inference needed to comprehend the passage will impede comprehension (Kintsch & Kintsch, 2005).

Reading comprehension difficulties may also emerge as a result of weaknesses in metacognitive awareness (e.g., a reader's knowledge and control over his or her thinking and learning; Baker & Brown, 1984). Readers with metacogntive awareness are able to monitor their understanding of the text's topic, source, and readability level during reading and apply strategic knowledge to improve comprehension (Medina & Pilonieta, 2006), whereas less skilled readers lack strategic knowledge and/or do not activate these skills during the reading process (Paris, Wasik, & Turner, 1996). In particular, readers with metacognitive awareness: (a) use words or imagery to elaborate content, (b) reread, paraphrase, or summarize text to clarify content, (c) deliberately link the text with prior knowledge and experience, and (d) identify areas of breakdown in comprehension and attempt to resolve the problem (Kintsch & Kintsch, 2005).

Interaction between reader and environment factors. The *sociocultural and instructional context* can impact reading comprehension (Rapp et al., 2007; Snow & Sweet, 2003). The experiences and opportunities that the family (e.g., Hart & Risley, 1995) and classroom teachers (e.g., Tharp & Gallimore, 1988) provide may vary as a function of the economic and cultural environment of the classroom, school, community, and larger society (Snow & Sweet). For example, differences in the value placed on reading, classroom instruction (e.g., direct instruction strategies have been shown to improve comprehension for struggling reader more effectively than less explicit instruction), economic resources (e.g., availability of texts, computers, and instructional materials), and caregiver's literacy levels may affect a reader's development of comprehension abilities (Rapp et al.; Snow & Sweet).

The *purpose* (e.g., internal versus external) and *consequence* (e.g., knowledge, application, and engagement) of a reading activity can also impact reading comprehension (Snow & Sweet, 2003). When reading is externally imposed (e.g., by a teacher, assignment, or assessment) versus internally imposed (e.g., entertainment or to obtain information), the reader's perception of the purpose of the task can have an impact on his or her attention, persistence, monitoring, and comprehension (Medina & Pilonieta, 2006). This is particularly salient for students with learning disabilities and/or attention disorders whose lack of task persistence can greatly impact their performance on externally imposed activities (Gersten et al., 2001). A reader's self-efficacy can also have a negative effect on their performance for externally imposed reading tasks (Caldwell, 2008). If the reader feels anxious and lacks confidence in his or her abilities, reading comprehension can suffer (Caldwell).

Rationale for Early Identification of Students who Struggle with Comprehension

As warned by the American Educator (1995), "if a child in a modern society like ours does not learn to read, he doesn't make it in life" (p. 3). Nearly 17.5 percent, roughly 10 million children in the United States will experience problems related to reading within their first three years of school (NRP, 2000). Research has shown that students who fail to learn how to read by the end of third grade are unable to catch up to peers (Simmons, Kuykendall, King, Cornachione, & Kame'enui, 2000). For example, through examining the literacy development of 54 first graders through their fourth grade year, Juel (1988) found that the probability of a first grader remaining a poor reader at the end of fourth grade was .88 and the probability of a first grader remaining an average reader at the end of fourth grade was .87. Similarly, in a large scale (*n* = 403) longitudinal study following children from kindergarten to ninth grade, Francis (1996) found that 74% of students identified with reading problems in third grade continued to have reading problems in ninth grade. In the United Kingdom, Cain and Oakhill (2006b) examined the profiles of 7- to 8-year-olds over a three-year period. Results revealed poor comprehenders maintained their status over the three year period (Cain & Oakhill). Taken together, these investigations suggest that without early identification and intervention students with poor reading skills will likely continue to perform below grade-level standards over time.

This notion is exemplified by Stanovich (1986) who applied the "Matthew Effect" (e.g., "the rich get richer and the poor get poorer") to reading deficits. Students who are strong achievers at the outset of school's subsequent academic progress is enhanced by experiences triggered from their initial success with reading, as a result, they read more which increases their word knowledge and helps them to become even better readers (Stanovich). By comparison, as a struggling reader attempts to read and experiences failure in accurately decoding words he or she will likely become frustrated and thus avoid reading, leading to a lack of practice in reading, no improvement, and a loss of self-efficacy and motivation for reading (Stanovich). This process increases the achievement gap between successful readers (e.g., those who have mastered the foundational reading skills) and struggling readers (e.g., those who have not mastered the foundational reading skills). Without intervention, reading problems will likely persist into adulthood. It is estimated that 93 million U.S. adults (age 16 and older) have basic (i.e., can perform simple and everyday literacy activities) or below basic (i.e., can perform no more than the most simple and concrete literacy tasks) literacy skills (Kutner et al., 2007).

Literacy skills can impact whether individuals receive high school diplomas or college degrees, their employment and earning potential, and community and civic involvement. Of the children who experience significant problems in reading, nearly 10 to 15 percent eventually drop out of high school and only 2 percent complete a four-year college program (Whitehurst & Massetti, 2004). Results of the National Assessment of Adult Literacy (2003) found that half of the adults who did not have high school diplomas performed at the below basic level (Kutner et al., 2007). Results also indicated that adults with higher literacy levels were more likely to be employed full-time and receive higher incomes as compared to adults with lower literacy levels who were more likely to be out of the labor force and generally earned lower incomes (Kutner et al.). Adults with higher literacy skills were found to be more likely to read to their children and discuss school topics, use the internet and e-mail, vote, volunteer, and access information about current local and national events (Kutner et al.). Children who experience significant problems in reading are also at risk for teen pregnancy, substance abuse, delinquency, and incarceration (Burke & Hagan-Burke, 2007; McGill-Franzen, 1987). According to Vanderstaay (2006), "the likelihood that a child will commit a delinquent act rises as school performance declines and falls as school performance improves" (p. 331). The National Assessment of Adult Literacy (2003) found that more than one million incarcerated adults in the nation had lower average literacy scores compared to adults in households (Kutner et al.).

Early identification of reading problems is critical to prevent the academic and social difficulties associated with reading deficits from persisting throughout school and into adulthood. Early identification has been recognized as an important preventive strategy across the fields of medicine, education, and mental health (Durlak, 1997; Elliott, Huai, & Roach, 2007). Alternative to the "wait-to-fail" model of identifying students in need of intervention, early identification, through systematic screening, provides opportunities for remediation prior to failure; therefore, reducing the incidence of academic and/or behavioral difficulties. Typically, screening measures are administered to all students in a grade, school, or district to identify individuals in need of additional support and provide opportunities for schools to respond to students' needs in order to improve academic and/or behavioral outcomes. The essential feature of a screening measure is its ability to discriminate students who are at risk for poor performance from those who are not at risk for poor performance (Jenkins, Hudson, & Johnson, 2007). Consequently, the challenge to effective remediation of reading deficits is identifying the right children at the right time (Torgesen, 1998). As noted by Foorman, Francis, Shaywitz, Shaywitz, and Fletcher (1997) the success rate of intervention drops significantly in later grades, with 82% of struggling readers becoming successful readers if intervention is provided in the early grades, 46% of children becoming effective readers if remediated at grades 3 to 5, and only 10 to 15% of children becoming successful readers if intervention is provided in later grades.

As a result of the No Child Left Behind Legislation (NCLB; 2002, PL 107–110), identifying students at risk for failure is crucial for schools. NCLB guidelines require schools to document adequate yearly progress through assessing academic outcomes (NCLB). In the area of reading, states typically accomplish this through administering a statewide reading assessment to students in designated grades (e.g., in Pennsylvania the statewide reading assessment is administered to students in Grades 3–8 and 11). Outcomes from the statewide reading assessment impact state and district-level decision making and policy, as a result they have important implications for individual students, teachers, schools, and districts (Shapiro, Solari, & Petscher, 2008).

Measurement of Reading Comprehension

Valid, reliable, and sensitive screening measures are needed to identify children at risk for reading problems, in particular students who are likely to perform poorly on statewide reading assessments. However, the ability to predict the children most at risk for reading comprehension problems has not been perfected, and the optimal combination of measures has not been determined (Bishop, 2003). This is likely due to the complex and multifaceted nature of reading comprehension, which makes it challenging to assess (RRSG, 2002). Therefore, it is not surprising that existing measures of comprehension differ vastly in (a) text type (narrative or expository), (b) reading format (silent or aloud), (c) time constraints (untimed or time limit per item or entire test), (d) level of measurement (word, sentence, or passage), (e) types of skills assessed, (f) response format (oral or written; forced-choice – true/false, yes/no, sentence verification; single-word; cloze/maze; multiple-choice; question and answer; open-ended; retell fluency; free/cued oral/written retell), (g) types of questions (literal, inferential, evaluative, or lexical), and (f) type of assessment (e.g., standardized, norm-referenced reading tests or curriculum-based measurement of reading ability) (Rathvon, 2004).

Standardized, norm-referenced reading comprehension tests. Standardized, norm-referenced reading comprehension tests assess overall comprehension skills and provide an indicator of student performance compared with same-age and grade-level peers. A wide variety of standardized, norm-referenced tests have been developed to assess reading comprehension. These measures differ in the degree to which they tap into text-based or situation-based comprehension, the nature of the comprehension task, and the response requirements (Berninger, Abbott, Vermeulen, & Fulton, 2006).

Typically, students read a brief narrative or expository passage and then answer literal or inferential comprehension questions about the setting, characters, plot, or sequence of events (narrative text) or about the main idea and supporting details (expository text). Standardized, norm-referenced reading tests are relatively easy to administer and score, with several measures available for group administration (e.g., California Achievement Test, Group Reading Assessment and Diagnostic Evaluation, Iowa Test of Basic Skills, Metropolitan Achievement Test, Stanford Achievement Test) or individual administration (e.g., Kaufman Test of Educational Achievement-Second Edition, Wechsler Individual Achievement Test-Second Edition, Woodcock-Johnson Third Edition Tests of Achievement, Woodcock-Johnson Third Edition Diagnostic Reading Battery, Woodcock Reading Mastery Test-Revised). Standardized, norm-referenced reading tests can be used to identify broad areas of strength and weakness for individuals so further appropriate action may be taken, as well as evaluate the effectiveness of school programs through measuring how groups of students are progressing in school (Caldwell, 2008). In particular, Farr (1999) described standardized, norm-referenced reading tests as "one important piece of information for planning, supporting, and evaluating school and system-wide curricula and instruction" (p. 52).

However, standardized, norm-referenced reading comprehension tests have been criticized for being content invalid, unlike real-life reading tasks, overly focused on lower-level lexical comprehension, indirectly measuring academic skills contained within a curriculum, ignoring the importance of fluency, lacking instructional utility, insensitive to change, and problematic for progress monitoring of students' growth over time (e.g., Barnett, Lentz, & Macmann, 2000; Elliott et al., 2007; Gresham & Witt, 1997; Klingner, 2004; Marston, 1989). In terms of content validity, research has demonstrated several problems with passage independence, which is an individual's ability to answer the test questions without reading the passage by using their world knowledge (Keenan & Betjemann, 2006). For example, Daneman and Hannon (2001), Katz, Blackburn, and Lautenschlager (1991), and Katz, Lautenschlager, Blackburn, and Harris (1990) demonstrated that students who were not permitted to read the passages on the Reading Comprehension subtest of the Scholastic Assessment Test achieved better than chance performance on the test questions. Likewise, Keenan and Betjemann found that more than half of the comprehension questions on the Gray Oral Reading Test (GORT; Wiederholt & Bryant, 1992, 2001) could be answered with above-chance accuracy without ever having read the passages. These findings raise concern as to whether these measures are truly assessing reading comprehension or whether they are merely assessing prior knowledge or verbal reasoning abilities (Keenan & Betjemann).

Placing students at the appropriate instructional level, using the test to document gain, and changing instruction and/or developing intervention plans from test results (i.e., instructional utility) have been identified as three misuses of standardized, norm-referenced reading comprehension tests (Royer & Lynch, 1983). Standardized, norm-referenced reading comprehension tests were not designed for documenting short-term changes in an individual's performance over time. Rather, they were designed to measure an individual's reading aptitude and predict future reading comprehension performance (Royer & Lynch). For these reasons, standardized, norm-referenced reading comprehension tests lack sensitivity to actual changes in an individual's performance.

Of the several criticisms of standardized, norm-referenced reading comprehension tests, lack of instructional utility is of primary concern. Assessment should inform intervention. It is difficult to use or link outcomes from standardized, norm-referenced tests to subsequent goal development and intervention planning (Macy, Bricker, & Squires, 2005). Typically, standardized, norm-referenced reading measures yield a performance score for the tested individual in relation to a predefined population (e.g., chronological age, gender, grade level). This score does not provide the necessary information required to create functional goals and an effective intervention plan to remediate reading deficits (Macy et al.). Standardized, norm-referenced reading comprehension tests often do not reflect an authentic picture of real-life reading tasks (i.e., how children respond in a familiar environment while involved in daily activities; Macy et al.). As a result, standardized, norm-referenced reading tests have been criticized for "focusing on what readers should be comprehending rather than what and how they are comprehending" (Klingner, 2004, p. 60). As noted by Kintsch and Kintsch (2005) current standardized, norm-referenced reading tests are "easy to use, but pay a heavy price for that" in that "some important comprehension skills do not come into play with such short texts, and deep understanding is not being assessed by the multiple-choice type questions used" (p. 87).

Curriculum-based measurement. As an alternative to standardized, norm-referenced tests, curriculum-based measurement (CBM) is designed for measuring students' academic proficiency in basic skill areas (i.e., reading, mathematics, written expression, and spelling) to determine when instructional modifications are necessary in general or special education and to monitor an individual's response to intervention (e.g., Christ & Silberglitt, 2007; Deno, 1985, 1986, 1989; Deno & Mirkin, 1977; Fuchs & Deno, 1991; Fuchs & Fuchs, 2002; Shinn, 1989, 1998, 2002). As highlighted by Deno (1985) and Marston (1989), curriculum-based measures are designed to be (a) reliable and valid measures of basic academic skills, (b) simple and efficient to administer, (c) easily understood by teachers, parents, and students, (d) inexpensive to produce in terms of time and resources, (e) vital signs of growth in basic skill areas, (f) sensitive to improvements in students' skills over time, (g) sensitive to the effects of intervention and short-term growth on an individual's skill level, (h) relevant to the content of instruction, (i) available in multiple forms of short duration to facilitate frequent administration, (j) based on production-type responses, so that a student's skills can be observed rather than inferred, and (k) relevant across a range of educational decisions.

Shinn (1998; 2002) described CBM as a dynamic indicator of skill development. CBM was considered "dynamic" because of its ability to detect short-term differences, as well as change over time, in an individual's skill level (Shinn, 1998; 2002). The use of frequently administered short assessments allows for systematic comparison of an individual's performance overtime. The term "indicator" was used to signify CBM as an empirically valid tool for measuring an individual's basic skills as an indicator of their overall performance in an academic area (Shinn, 1998; 2002). To illustrate this concept, Shinn (2002) described CBM as akin to a thermometer. Similar to a thermometer, CBM is a tool to assist in (a) making decisions about whether a problem warrants attention (e.g., thermometer: a temperature of a 101°F; CBM: performance score at the 25th percentile on a measure of reading ability), (b) determining the severity of the problem (e.g., thermometer: 101°F versus 105°F; CBM: performance score at the 25th percentile versus 10th percentile on a measure of reading ability), (c) setting goals for intervention (e.g., thermometer: return to 98.6°F; CBM: performance score at or above the 50th percentile on a measure of reading ability), (d) evaluating an individual's response to intervention, including the effectiveness of an intervention and an individual's progress towards a goal (e.g., thermometer: reduced temperature after Tylenol to 100°F; CBM: increased performance score to the 35th percentile after a repeated reading intervention to improve fluency), and (e) identifying when a problem has been remediated (e.g., thermometer: temperature consistently falls within the normal range; CBM: reading performance score consistently falls at or above the average range) (Shinn, 2002). Finally, CBM is a tool to assess "basic skills" in reading, mathematics, written expression, and spelling. CBM was not designed to assess broad content areas; rather, it was created to focus on the basic skills in the key subject areas (Shinn, 2002).

Oral Reading Fluency. Application of CBM for reading has predominantly examined the assessment of Oral Reading Fluency (ORF). ORF is the most widely used and thoroughly investigated CBM. ORF directly assesses the speed and accuracy of oral production of text (Adams, 1990; Berninger, Abbott, Billingsley, & Nagy, 2001). Students typically read aloud three short passages from grade-level controlled reading material. If the student hesitates on a word for 3-s, the examiner supplies the word and counts it as incorrect. The examiner also records errors of mispronunciation, substitution, omission, and transposition. Errors of insertion and repetition are ignored. Self-corrections provided within 3-s are counted as correct. The median number of words read correctly per one minute (WCPM) comprises the student's performance score.

Although it appears to be measuring only oral reading, research has provided strong support for ORF as an indicator of overall reading proficiency. For example, Fuchs, Fuchs, and Maxwell (1988) found strong correlations between ORF and a published measure of overall reading ability (i.e., Stanford Achievement Test; Gardner, Rudman, Karlsen, & Merwin, 1982, 1983). In particular, results indicated that ORF correlated higher with a criterion measure of reading comprehension (*r* = .92) than decoding (*r* = .81) (Fuchs et al.). This study was replicated by Shinn, Good, Knutson, Tilly, and Collins (1992) who also found higher correlations between ORF and measures of comprehension than measures of decoding. Numerous additional studies have established the validity of ORF as a measure of overall reading proficiency (see Dynamic Measurement Group [DMG], 2008; Shinn and Shinn, 2002). In a meta-analysis of 41 studies examining the correlations between ORF and standardized measures of reading achievement for students in grades 1 through 6, Reschly, Busch, Betts, Demo, and Long (2009), found correlations between ORF and measures of comprehension to range from .20 to .88, with most correlations greater than .51, yielding an average correlation of .65. One important finding from this study was that there were no significant differences found in the correlations between ORF and standardized measures of comprehension, decoding, and vocabulary, which provides additional support for ORF as a general outcome measure of overall reading ability (Reschly et al.).

Research has also demonstrated moderate to strong correlations between ORF scores and outcomes on state standardized reading assessments. At Grade 3, correlations between students' ORF scores and performance on the state reading assessment were .70 (Florida Comprehensive Assessment Test – Sunshine State Standards; Buck & Torgeson, 2003), .73 (North Carolina end of grade reading assessment; Barger, 2003), .74 (Arizona Instrument to Measure Standards; Wilson, 2005), and .79 (Illinois State Assessment Test; Sibley, Biwer, & Hesch, 2001). Correlations between third and fourth grade ORF scores and performance on the reading portion of the Ohio Proficiency Test ranged from .61 to .65 (Vander Meer, Lentz, & Stollar, 2005). At Grade 4, the relationship between students' ORF scores and performance on the reading portion of the Michigan Educational Assessment Program was .67 (McGlinchey & Hixson, 2004). Also at fourth grade, correlations between Grade 4 performance on the reading portion of the Washington Assessment of Student Learning and fall, winter, and spring ORF scores were .50, .51, and .51 respectively (Stage & Jacobsen, 2001). Correlations between students' ORF scores and performance on the reading portion of the Colorado State Assessment Program ranged from .73 to .80 in third grade (Shaw & Shaw, 2002), .67 in fourth grade, and .75 in fifth grade (Wood, 2006). Finally, in Pennsylvania, correlations between students' ORF scores and performance on the reading portion of the Pennsylvania System of School Assessment were .69 and .71 in second grade (Keller-Margulis, Shapiro, & Hintze, 2008), ranged from .65 to .68 in third grade (Shapiro, Keller, Edwards, Lutz, & Hintze, 2006; Shapiro et al., 2008), ranged from .64 to .69 in fourth grade (Keller-Margulis et al.; Shapiro et al., 2008), and ranged from .62 to .75 in fifth grade (Shapiro et al., 2006; Shapiro et al., 2008). Additional investigations of the relationship between students' ORF scores and performance on the reading portion of the state assessment have been conducted in Florida (Castillo, Torgeson, Powell-Smith, & Al Otaiba, 2003; Roehrig, Petscher, Nettles, Hudson, & Torgeson, 2008), Iowa (Fuchs, Fuchs, Hosp, & Jenkins, 2001), Minnesota (Hintze & Silberglitt, 2005; Silberglitt, Burns, Madyun, & Lail, 2006; Wiley & Deno, 2005), and Oregon (Crawford, Tindal, & Stieber, 2001; Good, Simmons, & Kame'enui, 2001).

The essential feature of a screening measure is its ability to predict students who are at risk for poor performance from those who are not at risk for poor performance (Jenkins et al., 2007). Research has consistently found ORF scores to be an effective tool for identifying students in need of additional reading instruction (Fuchs & Fuchs, 1992; Jenkins & Jewell, 1993). In particular, research has found ORF scores to be a significant predictor of students' performance on state standardized reading assessments. On the Dynamic Indicators of Basic Early Literacy Skills (DIBELS, 6th edition; Good & Kaminski, 2002), at third grade, 110 WCPM and higher has been identified as an appropriate cut score for predicting success (i.e., low risk) on state reading assessments, whereas scores at 80 to 109 WCPM are considered at some risk and scores below 80 WCPM are at high risk for not meeting grade-level expectations on state reading assessments (Good, Simmons, Kame'enui, & Wallin, 2002). For example, of the third grade students who were classified as low risk based on earning an ORF score greater than or equal to 110 WCPM, 81.9% passed the reading portion of the Arizona Instrument to Measure Standards (AIMS; Wilson, 2005). Likewise, of the third grade students who were classified as at risk based on earning an ORF score less than 80 WCPM, 93% did not demonstrate proficiency on the reading portion of the AIMS (Wilson). Comparable results were found in other states; of the third grade students who were identified as low risk based on earning an ORF score greater than or equal to 110 WCPM, 90% (Colorado State Assessment Program; Shaw & Shaw, 2002), 91% (Florida Comprehensive Assessment Test – Sunshine State Standards; Buck & Torgeson, 2003), and 99% (Illinois

State Assessment Test; Sibley et al., 2001) were successful on their respective statewide reading assessment.

Research has also examined the diagnostic accuracy of ORF to predict performance on state standardized reading assessments, which includes a measure's sensitivity (i.e., its accuracy in identifying students at risk for not meeting grade-level expectations on the statewide reading assessments) and specificity (i.e., its accuracy in identifying students not at risk). For example, Shapiro et al. (2008) examined the diagnostic accuracy of fall and winter ORF scores and 4Sight Reading Benchmark Assessment scores (4Sight; Success for All Foundation, 2007) for predicting performance on the reading portion of the Pennsylvania System of School Assessment (PSSA; Data Recognition Corporation [DRC], 2007) for students in grades three through five. Results indicated excellent sensitivity and weak specificity for ORF scores across grades with the combination of ORF and 4Sight resulting in better classification rates compared to ORF or 4Sight alone (Shapiro et al.). The sensitivity was .95 and .96 in third grade, .79 and .88 in fourth grade, and .93 and .97 in fifth grade, whereas the specificity was .55 and .59 in third grade, .49 and .59 in fourth grade, and .58 and .61 in fifth grade (Shapiro et al.). Conversely, using 110 WCPM as the cut score for third grade performance on the reading portion of the Florida Comprehensive Assessment Test, Buck and Torgeson (2003) found weaker sensitivity (.77) and greater specificity (.92). Taken together, these findings indicate that while ORF is not a perfect metric, it has utility as a screening tool to identify those students at risk for poor performance on state reading assessments.

Despite strong empirical support for ORF as an indicator of students' overall reading proficiency, research has found that ORF is not a strong indicator of reading comprehension as students advance in grade level; by fourth and fifth grade, ORF may not offer as much sensitivity for measuring students' reading comprehension skills as other measures (Hosp & Fuchs, 2005; Shapiro et al., 2008; Shinn et al., 1992). Hosp and Fuchs assessed whether the relation between ORF and specific reading skills changed as a function of grade level. Similar to Shinn et al., who reported a stronger relationship between ORF and measures of word reading ability in third grade over fifth grade, Hosp and Fuchs reported stronger relations between ORF and a criterion measure of word reading (i.e., Woodcock Reading Mastery Test-Revised; Woodcock, 1987) at Grades 1, 2, and 3 compared to Grade 4. In addition, stronger relations were found between ORF and a criterion measure of decoding (i.e., Woodcock Reading Mastery Test-Revised) for second and third grades over fourth grade (Hosp & Fuchs). Jenkins and Jewell (1993) examined the relationship among the performance of 335 students from Grades 2 to 6 on ORF and the total reading and comprehension subtests of the Gates-MacGinitie Reading Tests (MacGinitie et al., 1978) and the Metropolitan Achievement Tests (MAT-6; Prescott et al., 1984). Results indicated that the relationship between ORF and the criterion measures of reading decreased from Grades 2 to 6 (Gates-MacGinitie, Total Reading: Grade 2 r = .83 to Grade 6 r = .67; Comprehension: Grade 2 r = .86 to Grade 6 r = .63; MAT-6, Total Reading: Grade 2 r = .87 to Grade 6 r = .60; Comprehension: Grade 2 r = .84 to Grade 6 r = .58) (Jenkins & Jewell). Kranzler, Miller, and Jordan (1999) also found correlations between ORF and a criterion measure of comprehension to decrease with grade (California Achievement Test, Comprehension: Grade 2 r = .63 to Grade 5 r = .51).

Whereas in early elementary ORF alone has been identified as a strong indicator of overall reading ability, Speece et al. (2010) found that the overall reading ability of the fourth grade students in their sample could not be defined by a single metric. Rather, three factors emerged (comprehension, word reading, and fluency) to explain these students' reading competence. This finding provides support for a multivariate approach to reading ability screening. The goal of any assessment is to "maximize the representativeness of the content sampled for the test, making it possible to generalize results to the content domain" (Campbell, 2005, p. 347). The predictive power of ORF appears to decrease as grade level increases. Therefore, direct measurement of comprehension is needed, especially for students in higher grade levels. Adding a measure of reading comprehension to ORF will likely enhance the decision making process for identifying proficient and non-proficient readers.

An additional reason for including a direct measure of reading comprehension is to increase the face and content validity of the assessment. Despite strong empirical support for ORF as an indicator of students' overall reading proficiency, a primary barrier to acceptance among teachers of ORF as an index of overall reading proficiency is its lack of face validity as a measure of reading comprehension (Roberts, Good, & Corcoran, 2005; Shinn et al., 1992; Williams, Skinner, Floyd, Hale, Neddenriep, & Kirk, 2011). ORF requires students to read text aloud and does not *directly* assess students' understanding of what they read; consequently, teachers may not view ORF as having the needed face validity to accept it as a measure of reading comprehension or content validity to design interventions for students who struggle with comprehension. As noted by Fewster and MacMillan (2002) providing teachers with information regarding the technical adequacy of ORF does not necessarily improve its face validity. In particular, practitioners report concern with ORF's ability to detect "word callers" (Dewitz & Dewitz, 2003; Hamilton & Shinn, 2003; Meisinger, Bradley, Schwanenflugel, Kuhn, & Morris, 2009; Roberts et al.; Shapiro, 2004). "Word callers" are readers who have fluent decoding without high levels of comprehension (Stanovich, 1986). Adding a measure of comprehension to ORF will likely enhance the decision making process for detecting "word callers" and inform intervention development by assisting in identification of specific comprehension skill deficits.

Free oral retell. One potential method for enhancing ORF's measurement of reading comprehension is Free Oral Retell. Free oral retell requires an individual to read a passage and then retell the key information from the passage in their own words. The term "free" indicates that the oral retell is not prompted, such that the examiner does not provide cues to assist the individual in retelling the text. Free oral retell involves the coordination between a reader's past experiences, knowledge of the topic, familiarity with text structure, and knowledge of language (Copmann & Griffith, 1994; Hansen, 1978). Compared to other methods of measuring reading comprehension, a primary advantage of free oral retell is that it measures a broad range of comprehension skills that are directly linked to instruction and intervention (Klingner, 2004; Roberts et al., 2005). Retelling assists in determining a reader's understanding of text (Ringler & Weber, 1984) through providing information about a reader's process for remembering and sequencing the events and major concepts presented in the text (Hansen). Retelling provides a view of the quantity, quality, and organization of information a reader amassed from the passage (Winograd, Wixson, & Lipson, 1989). Taken together, retelling meets the criteria outlined by the RRSG (2002) for reading comprehension measures, including providing information regarding a reader’s knowledge (i.e., their understanding of text), application (i.e., their ability to tell about what they read in their own words), and engagement (i.e., their process for remembering and sequencing text and the quantity and quality of information amassed).

Retelling is the “most straightforward assessment” of a reader’s interaction with the text (Johnson, 1983, p. 54). Roberts et al. (2005) highlighted several benefits of retelling over other comprehension metrics including: (a) yielding a large sample of comprehension behaviors that can inform intervention and increase the chance of detecting post-intervention change, (b) time efficiency in creating, administering, and scoring the assessment, and (c) greater ease in informing instruction. For example, using the cloze technique in combination with ORF requires students to read a separate passage from the one used for ORF, thus decreasing the efficiency of the assessment. Likewise, question-response tests (e.g., multiple-choice) are limited in terms of efficiency as it can be difficult to write good comprehension questions and create parallel tests, especially when the varying levels of background knowledge of students is taken into account. In particular, response options must be chosen with care so that the nature of the distracters does not mislead or guide the participants to the correct response (i.e., passage independence) (Cain & Oakhill, 2006a). Sentence verification tests are also limited in terms of efficiency as a large number of items are required for task sensitivity (Cain & Oakhill). Specifically, Royer (1990) indicated that six passages (96 sentences) are required to obtain reliabilities between .8 and .9.

Furthermore, compared to free oral retell, several of these measures yield a limited sample of behaviors. For example, high levels of performance on cloze tasks may be obtained if participants have good local sentence-processing skills (Cain & Oakhill, 2006a). In particular, Shanahan, Kamil, and Tobin (1982) found participants perform fairly well on cloze passages in which the sentences have been scrambled, indicating that cloze techniques are measuring sentence- rather than text-processing skills. Consequently, cloze assessments may fail to detect children with reading comprehension difficulties. Forced choice (e.g., true/false or yes/no) and sentence verification tests may be good indicators of memory for literal details in the text, but are limited in their assessment of specific comprehension skills, such as inference making (Cain & Oakhill). Poor comprehenders often have specific problems generating inferences (Taylor, Alber, & Walker, 2002). Recognizing that a statement is consistent with the text representation is not the same as generating inferences (Cain & Oakhill), thus forced choice and sentence verification tests may fail to identify children with comprehension problems.

An additional advantage of retelling is that it can be taught, modeled, and practiced more easily than the other methods of assessing comprehension. In particular, retelling demonstrates consequential validity, that is, it has a positive consequence for the individual as a result of the experience (McKenna & Stahl, 2009). Retelling may improve a reader's connection with the text including processing of story structure and identification of important elements of the text (McKenna & Stahl). For instance, Gambrell, Pfeiffer, and Wilson (1985) examined the impact of practice on students' retell, as well as differences between students' oral retell and illustration of important details. Results indicated that practice in retelling generalized to different texts and resulted in greater recall for the oral retell group than illustration group (Gambrell et al.). Students in the oral retell group remembered more important ideas from the story, made greater elaborations, and answered more literal and inferential comprehension questions correctly than the illustration group (Gambrell et al.).

Effective retelling consists of a coherent structure (e.g., organized and succinct), shows consideration for the text genre (e.g., narrative versus expository text structure), and draws on prior knowledge and past experiences (Copmann & Griffith, 1994). The sequence in which a student retells the text provides important information about how he or she prioritizes, chunks, synthesizes, and expresses information without prompting or cueing (Blackowicz & Ogle, 2001). Unfamiliarity with text structure or content can impede a student's retell. Meyer, Brandt, and Bluth (1980) found that readers who lack awareness of text structure often do not approach reading with a plan, which leads to inaccurate chunking of information and poor retrieval. According to Caldwell and Leslie (2005) a good narrative retell includes the major story elements (e.g., characters, goal/problem, events, resolution), is sequential, and makes causal connections between events in the story, whereas a good expository retell is guided by knowledge of the topic and expository text structure, is retold in a sequential or time-ordered format, identifies important information (e.g., main idea and details), and may include cause and effect, problem and solution, or compare and contrast.

Appendix A provides an overview of the research that has been conducted using oral retell to assess reading comprehension. Studies on retelling published in journals in the fields of education and social sciences were identified using computer-based searches in PsychInfo, ERIC, *School Psychology Review*, *Journal of School Psychology*, *Reading*

Research Quarterly, *Reading and Writing Quarterly*, *Reading Research and Instruction*, *The Reading Teacher*, *Reading*, *Reading Today*, and *Journal of Adolescent and Adult Literacy*. References cited in identified studies were also searched for relevant information and citations of other relevant studies. The following search terms were used: *retelling, retell, recalling, recall*, *oral retell*, *free oral retell*, *retell fluency*, *text analysis*, *propositions*, *idea units*, *rubric*, and *reading comprehension assessment* or *measurement*. Only studies which utilized school-aged samples and in which the participants read the passage were included in this review. Several additional studies have been conducted with school-aged children in which the examiner read the text to the participants (e.g., Copmann & Griffith, 1994; Gardill & Jitendra, 1999; Glenn, 1980; Morrow, 1985; Morrow, Sisco, & Smith, 1992; Moss, 1997; Vosniadou & Schommer, 1988); these studies were excluded from the table in Appendix A since they examine listening comprehension more than reading comprehension.

Retelling has frequently been used as an assessment tool of reading comprehension in reading research (Fuchs et al., 1988; Gambrell et al., 1985; Johnson, 1983); however, it has less frequently been used for assessment in applied settings (Blachowicz & Ogle, 2008; Maria, 1990). This is likely due to the scoring of retells. Whereas retelling is a "straightforward and feasible assessment strategy in terms of initial preparation (requiring only selection of suitable material), methods of scoring recalls can be extremely difficult and time consuming to implement" (Fuchs et al., p. 21). Typically, researchers use a text analysis system to divide the passage into idea units (i.e., propositions) and assign idea units a particular level of importance. The oral retell is transcribed with two independent scorers determining the number of idea units contained in the oral retell. Of the 23 studies identified in Appendix A, more than half utilized a text analysis system for scoring the oral retell.

For example, Best et al. (2008) employed a text analysis system to score participants' free oral retell by dividing each sentence (i.e., idea unit) of the story into main propositions (main ideas) and subpropositions (supporting details for each main idea). This technique resulted in 61 main propositions and 43 subpropositions for the narrative text and 45 main propositions and 47 subpropositions for the expository text (Best et al.). Each participant's oral retell was transcribed and broken into idea units (subject, verb, and object), with each idea unit scored against the identified main propositions and subpropositions for the narrative or expository text. Participants received 1 point for each idea unit that matched the main proposition and provided more detailed information in the subproposition. Participants received 0.5 points for each idea unit that matched the main proposition but did not provide any information about the subproposition. The participant could also receive 0.5 points for each idea unit that matched the main proposition, but provided erroneous information. Zero points were awarded for each idea unit that did not match any proposition (Best et al.). Using the sentence "Plants need sunlight, water, and air to live" as an example, a 1 point response may include "plants need water, sunlight and air" and a 0.5 point response may include "plants need water to live" or "plants do not need sunlight" (Best et al., p. 146). The total points a participant earned for free recall of a passage was divided by the total number of propositions in each passage.

This method is not realistic or feasible for school settings; it is unlikely that a teacher would have enough training in using the text analysis scoring system, time to develop the scoring template, and time to have two independent teachers transcribe and score each student's oral retell (Maria, 1990). Alternative to a text analysis scoring system, researchers have primarily chosen to (a) count the total number of words included in a participant's retell (e.g., Burke & Hagan-Burke et al., 2007; Fuchs et al., 1988; Marcotte & Hintze, 2009; Pressley, Hilden, & Shankland, 2005; Riedel, 2007; Roberts et al., 2005) or (b) examine the story structure elements included in a participant's retell (e.g., Gambrell, Koskinen, & Kapinus, 1991; Hagtvet, 2003; Rabren, Darch, & Eaves, 1999; Shannon, Kame'enui, & Baumann, 1988; Short, Yeates, & Feagans, 1992).

Counting the total number of words included in a participant's retell. In order to identify the most feasible method for scoring oral retells, Fuchs et al. (1988) examined the following scoring methods: (a) counting the total number of words retold, (b) calculating the percentage of content words retold (e.g., proper nouns, common nouns, verbs, adjectives, or adverbs that exactly match or are synonyms for words in the passage), or (c) computing the total number of idea units retold (e.g., subject, verb, and object). In this study, participants' were allotted 5 minutes to read a 400-word folktale passage and 10 minutes to retell the passage. Participants' retells were audio recorded and transcribed for scoring. Results from their study revealed that different methods of scoring related comparably with each other (*r*= .84. .86, and .94) and had similar correlations with the other measures of reading ability (Fuchs et al., 1988), with the most feasible scoring method being counting the total number of words retold (Fuchs et al., 1988). In particular, the three measures of oral retell demonstrated the highest correlations with the total number of words read correctly (*r* = .63, 73, and .65) and average number of questions answered correctly (*r* = .65, .70, and .64). Moderate correlations were also found between the three oral retell scoring methods and the three written retell scoring methods (*r* = .58 to .65), the Reading Comprehension subtest of the Stanford Achievement Test (SAT-7; Gardner et al., 1982, 1983) (*r* =.59, .64, and .60), three methods of scoring written cloze (*r* = .50 to .63), and three methods of scoring oral cloze (*r* = .43 to .61). Despite the feasibility over the other scoring methods, counting the total number of words retold required transcription of the student's retell and multiple scorers; therefore, decreasing the feasibility of this method for the classroom setting.

Alternatively, Good and Kaminski (2002) as part of the DIBELS 6th edition and Roberts et al. (2005) as part of the Vital Indicators of Progress (VIP) used within the Voyager Universal Literacy System have developed a method for counting the total number of words included in a participant's retell that does not require transcription or use of multiple scorers. As part of the ORF assessment, participants are allotted 1 minute to read a 200 or more word grade-level controlled passage. After the 1 minute time period, the passage is removed, and the participant is given 1 minute to retell what they just read in their own words. As the participant is orally responding, the examiner records the total number of words that the participant can retell within a one minute time period. The Retell Fluency (RTF) metric was created to complement ORF, as a means of improving ORF's face validity (Good & Kaminski). It was designed to prevent children from focusing on fluency without attending to meaning (i.e., speed-reading) and identify children whose comprehension is inconsistent with their ORF (i.e., "word callers") (Good & Kaminski). In particular, Good and Kaminski suggest that:

> a rough rule of thumb may be that, for children whose retell is about 50% of their oral reading fluency score, their oral reading fluency score provides a good overall indication of their reading proficiency, including comprehension. But, for children who are reading over 40 words per minute and whose retell score is 25% or less of their oral reading fluency, their oral reading fluency score alone may not be providing a good indication of their overall reading proficiency. (p. 31)

The DIBELS 6th edition RTF measure is limited in three ways. First, there are no scoring guidelines for the exact number of words that should be included in an effective retell, thus limiting the instructional utility of the DIBELS 6th edition RTF measure. Although, the "rough rule of thumb" may be helpful in identifying students at-risk for comprehension problems, without concrete scoring information, it is unclear how this measure can be used for instructional decision-making and progress monitoring. Second, the DIBELS 6th edition RTF measure alone does not account for the quality of the retell. For example, "mistakes or inconsistencies in the retell do not count against the student as long as the student is still on topic" (Good & Kaminski, 2002). A student's retell may include a lot of words, placing them in the low risk range; despite limited and/or scrambled ideas and inaccuracies in their retell. Conversely, a student's retell may concisely summarize all of the key story structure elements, placing them in the at risk range by reason of their retell including fewer words. Therefore, the DIBELS 6th edition RTF measure alone may not be an accurate indicator of a reader's overall comprehension abilities. Furthermore, as noted by Bellinger and DiPerna (2011), the DIBELS 6th edition RTF measure "may be an insufficient measure of comprehension" because the assessment is based on a "short period of time" (i.e., 1-minute oral reading task followed by a 1-minute retell task), thus limiting the “amount of meaningful information that the child could comprehend” (p. 418).

Third, there is limited and conflicting research on the psychometric properties of the DIBELS 6th edition RTF measure. Technical Adequacy information for the DIBELS 6th edition RTF measure yields more consistent and higher correlation coefficients for ORF WCPM compared to RTF score for (a) reliability of single probe (WCPM across 4 studies spanning grades 1 through 6 mean r = .92; range, r = .83 to .98; RTF 1 study in grade 1 r = .57) and multi-probe (WCPM across 4 studies spanning grades 1 through 6 mean r = .97; range, r = .94 to .99; RTF 1 study in grade 1 r = .87), (b) concurrent criterion-related validity with variety of measures (e.g., GRADE; Williams, 2001; TerraNova, CTB/McGraw-Hill, 2002) (WCPM across 20 studies spanning grades 1 through 6 mean r = .70; range, r = .42 to .97; RTF across 4 studies in grades 1 and 3 mean r = .58; range, r = .42 to .81), and (c) predictive criterion–related validity with variety of measures (e.g., GRADE and TerraNova) (WCPM across 14 studies spanning grades 1 through 6 mean r = .70; range, r = .29 to .94; RTF across 1 study in grade 1 mean r = .42; range, r = .39 to .46) (DMG, 2011b).

Burke and Hagan-Burke (2007) examined the relationship between RTF and the other DIBELS 6th edition measures. RTF had low to moderate correlations with the other DIBELS measures (r = .26 to .69), having the strongest relation with ORF. This is not surprising given that a student’s performance on RTF is dependent on his or her performance on ORF. Whereas ORF emerged as the single strongest predictor of the Test of Word Reading Efficiency Phonemic Decoding Efficiency (PDE) and Sight Word Efficacy (SWE) subtests, RTF did not explain any variance in PDE or SWE after controlling for DIBELS Phoneme Segmentation Fluency, Nonsense Word Fluency, and ORF.

Marcotte and Hintze (2009) examined the incremental and concurrent validity, as well as the predictive utility, of the DIBELS 6th edition RTF measure compared to other formative measures of reading comprehension. Results revealed that RTF consistently showed the weakest relationship with the other measures (r = .45 to .49), having the strongest relation with ORF (Marcotte & Hintze). By comparison, sentence verification technique (SVT) and written retell (WRT) correlated with the other measures between .49 and .59, ORF correlated with the other measures between .56 and .72, and maze (MZ) correlated with the other measures between .57 and .72 (Marcotte & Hintze). Furthermore, results from the regression analysis indicated that RTF did not contribute to the prediction of the GRADE; it was the weakest predictor variable and the only non-significant variable (Marcotte & Hintze). By comparison, ORF was the strongest predictor, followed by MZ, SVT, and WRT (Marcotte & Hintze). Finally, low levels of interscorer agreement were reported for the RTF measure. Marcotte and Hintze found that, using the 2-point criterion for judging interscorer agreement set by Good and Kaminski (2002), only 33% of the interscorer checks on the RTF were within 2-points of one another, with 46% of the interscorer checks within 3-points of one another. The range of interscorer agreement was between 0 to 15 words (Marcotte & Hintze).

Riedel (2007) also investigated the concurrent validity and predictive utility of the DIBELS 6th edition RTF measure with the GRADE. Similar to Marcotte and Hintze (2009), who found RTF and ORF to correlate with the GRADE .46 and .65 respectively, Riedel found RTF to correlate with the GRADE .41 and .51 and ORF to correlate with the GRADE .59 and .67. RTF was also found to be a weaker predictor of comprehension than ORF. In addition, RTF did not substantially improve the predictive accuracy compared to ORF alone. Alone, ORF yielded classification accuracies of 67.9% and 71.8% (Riedel). In both cases, RTF added minimally (0.8% and 0.6% respectively) to ORF's prediction of the GRADE (Riedel). Overall, Riedel concluded that there is a "lack of empirical evidence for the usefulness of the RTF task" (p. 560).

As noted by Bellinger and DiPerna (2011), the scoring procedures for the DIBELS 6th edition RTF measure "has the potential be a challenging and possibly unreliable practice" perhaps because "student's speech may be faster than an examiner can accurately count" (p. 418). Pressley, Hilden, and Shankland (2005) evaluated the reliability of the DIBELS 6th edition RTF measure by comparing scores obtained from the live scoring of the RTF with scores obtained from transcribed re-scoring of third grade students' audio-recorded responses. Similar to Marcotte and Hintze (2009) who found low levels of interscorer agreement, Pressley et al. found significant differences between the live RTF scoring and transcribed re-scoring of the total number of words retold (mean difference of 11 words). The inaccuracy in the live RTF score had a large effect size (mean = .95; range = .89 to 1.00; Pressley et al.). Replication and extension of the Pressley et al. study with fourth grade students by Bellinger and DiPerna yielded similar findings with a significant difference found between real time RTF scores and recorded RTF scores for each passage (mean difference of 32 words) and a large effect size for each of the passages ($3.83 \leq$ Cohen's $d \leq 4.12$). These findings suggest that the DIBELS 6th edition RTF measure lacks adequate reliability for scoring students' retell.

Pressley et al. (2005) also evaluated the validity of the DIBELS 6th edition RTF measure compared with propositional analysis. On average, participants included 42 words in their retell within the one-minute period (Pressley et al.). Comparison of the total propositions in each story against the actual propositions retold indicated that students' retellings of idea units were low, and students included very few of the propositions in their story retells that contributed to their RTF score (Pressley et al.). For example, the "Pots" story had a total of 85 idea units for the entire passage; on average, participants included seven idea units in their retell (Pressley et al.). Taken together, these findings caused the researchers to "wonder whether the DIBELS retelling data are of any value whatsoever" (p. 25). Of primary concern is the weak instructional utility of the measure. Pressley et al. noted that:

> Recall of individual words and counting of individual words, as the DIBELS calls for, conceptually makes no sense based on what is known about the comprehension of text, with comprehension of ideas and relationship between ideas being what matters more than individual concepts or words in the text. The DIBELS as currently specified does not assess understanding or memory of ideas. (p. 25–26)

Consequently, the DIBELS 6th edition RTF score does not possess the advantages of free oral retell previously outlined. RTF does not provide information about the quality and organization of information a reader amassed from text, nor does it yield a large sample of comprehension behaviors that can inform instruction and intervention.

Examining the story structure elements included in a participant's retell. A potentially better method for judging an effective retell may include evaluating the accuracy of the components, sequence, and coherence of the retell (Caldwell & Leslie, 2005). This can be accomplished through the use of a rubric, in which students are awarded points for each idea or fact recalled (Blachowicz & Ogle, 2001). The rubric method of scoring allows for greater conceptual match to what we know are the key elements of reading comprehension, in particular a reader's comprehension of the story structure and sequence of information recalled.

Five studies have used story structure elements as the methodology for scoring a participant's retell (see Appendix A); however, none of these studies have investigated the psychometric properties of using story structure elements to score oral retell. In addition, each of these studies utilized a different methodology for scoring the story structure elements, including (a) counting the total number or proportion of story elements included in the retell (e.g., setting, theme, plot, and resolution; Gambrell et al., 1991), (b) awarding varying point amounts for including specific story elements in the retell (e.g., introduction ≤ 3 points, cause/motive ≤ 3 points, or event 1 ≤ 2 points; Hagtvet, 2003), (c) awarding points on a sliding scale based on the amount of information provided for the specific story element in the retell (e.g., 0 = no mention of setting, 1 = vague representation of setting, 2 = accurate representation of setting, or 3 = verbatim representation of setting; Short et al., 1992; 0 = no detail-cues included in retell, 0.5 = one detail-cue included in retell, or 1 = two detail-cues included in retell; Shannon et al., 1988), and (d) scoring the quality of information provided for the specific story element in the retell (e.g., none, low, moderate, or high degree; Rabren et al., 1999).

The DMG (2011a) recently released a revised version of the DIBELS (i.e., DIBELS Next). The DIBELS Next includes a revised version of the RTF measure called

DIBELS Oral Reading Fluency Retell (DORF Retell). The DORF Retell continues to provide information regarding the quantity of the retell. However, the directions are slightly different. Instead of counting the number of words the student produces, the examiner is only to count the number of words that the student says that are *related to the passage*. It is noted that “the assessor must make a judgment about the relevance of the retell to the story while drawing the line” to record the number of words related to the story retold (DMG, 2011b, p. 26). Independent research examination of the reliability of the scoring changes made to the DORF Retell has yet to occur. Previous research on the DIBELS 6th edition RTF measure yielded low levels of interscorer reliability and significant differences between the real time RTF scores and recorded RTF scores (e.g., Bellinger & DiPerna, 2011; Marcotte & Hintze, 2009; Pressley et al., 2005). It is conjectured that the increased subjectivity and complexity of the new scoring procedure for the DORF Retell will result in similar or lower levels of interscorer reliability and significant differences between the real time RTF scores and recorded RTF scores.

Despite potential scoring issues, conceptually, the DORF Retell is an improvement on the DIBELS 6th edition RTF measure because it also provides information regarding the quality and sequence of information recalled. The quality and sequence of information recalled are assessed through the newly added (a) “Quality of Response” rating, which requires the examiner to indicate from 1 to 4 how many details related to the main idea were provided in the retell and whether the details were provided in a meaningful sequence (i.e., 1 = *provides 2 or fewer details*; 2 = *provides 3 or more details*; 3 = *provides 3 or more details in a meaningful sequence*; or 4 = *provides 3 or more details in a meaningful sequence that captures a main idea*) and (b) “General Retell

Response Patterns" checklist, which requires the examiner to record whether across all three passages the participant: *summarizes*, *repeats the same detail*, *retells the passage verbatim*, *"speed reads" the passage and has limited retell relative to number of words read*, or *talks about own life related to passage* (DMG, 2011a). In addition, the DORF Retell now includes benchmark goals and cut points for (a) the exact number of words *related to the passage* that should be included in an effective retell and (b) the quality of response rating that reflects a student's qualitative understanding of the passage, thus allowing for interpretation of the DORF Retell score. Note, the DORF Retell was released after this dissertation research study was conducted.

Consistent with research on the technical adequacy of the DIBELS 6th edition RTF measure, in the technical manual for the DIBELS Next, the creators of the DORF Retell measure report more consistent and higher correlation coefficients for DORF WCPM compared to DORF Retell for alternate-form reliability (WCPM range, *r* = .92 to .98; Retell range, *r* = .65 to .81), test-retest reliability (WCPM range, *r* = .91 to .97; Retell range, *r* = .27 to .69), concurrent validity with the GRADE (WCPM range, *r* = .61 to .75; Retell range, *r* = .40 to .65), and predictive validity with the GRADE (WCPM range, *r* = .59 to .77; Retell range, *r* = .48 to .61) (DMG, 2011b; Powell-Smith, Good, Latimer, Dewey, & Kaminski, 2011).

Reed (2011) examined the psychometric properties of the DORF Retell along with 10 other commercially or publically available retell measures. All but one measure (i.e., VIP) examined the story ideas or facts recalled. Reed concluded that all of the retell measures reviewed provided insufficient information regarding the psychometric properties of the instruments resulting in a lack of confidence in the existing retell measures' ability to assess students' reading comprehension and inform intervention development. In particular, Reed indicated that future research should seek to improve both the technical adequacy and practical relevance of retell measurement in order to possess instructional utility.

Reading Retell Rubric

Given the limited and varied research on examining the story structure elements included in a participant's retell, future research was warranted to examine the technical adequacy and utility of using a rubric method for scoring oral retell. The Reading Retell Rubric (RRR) for narrative text and expository text were developed through a review of the literature and examination of story elements that could be identified in commercially available oral reading probes. For example, according to Caldwell and Leslie (2005) a good narrative retell includes the major story elements (e.g., characters, goal/problem, events, resolution), is sequential, and makes causal connections between events in the story, whereas a good expository retell is guided by knowledge of the topic and expository text structure, is retold in a sequential or time-ordered format, identifies important information (e.g., main idea and details), and may include cause and effect, problem and solution, or compare and contrast. Medina and Pilonieta (2006) also described the differences between narrative (e.g., character, plot, temporal sequence, often past tense) and expository text (e.g., informational, includes technical vocabulary, does not necessarily follow a timeline) and highlighted key aspects that should be included in a narrative retell, such as characters, setting – place and time, problem, sequence of actions, and resolution to the problem.

Narrative and expository texts differ in person (e.g., narrative texts are generally about people or characters and written from a personal perspective), orientation (e.g., expository texts are subject-oriented), time (e.g., narrative texts link events in a chronological order), and linkages (e.g., expository texts link events in a logical order) (Copmann & Griffith, 1994); consequently, the RRR for narrative text consists of different its than the RRR for expository text. The narrative version of the RRR was designed to measure a student's ability to retell the following story structure elements: (1) *Theme*: the central idea or point of the passage; (2) *Problem*: an obstacle or conflict the main character must resolve; (3) *Goal*: how the main character wants the problem to be resolved or what the main character is attempting to achieve; (4) *Setting*: where and when the story takes place; (5) *Characters:* people or animals in the story; (6) *Initiating event:* an idea or action that sets further events in motion or causes the main character to respond in some way; (7) *Climax*: when the conflict or problem is resolved; (8) *Sequence*: retells the story in a structural or temporal order; (9) *Problem solution:* how the problem was resolved; and (10) *End of story:* conclusion or how the story turns out. Students could earn up to 10 points, 1 point for recalling each of the story structure elements. The expository version of the RRR was designed to measure a student's ability to retell the following story structure elements: (1) *Topic:* the subject of the text; (2) *Main idea*: what the text is all about or most of the sentences are about or the overarching theme; (3) *Primary supporting details*: facts needed to understand the main idea or support the main idea by explaining it and developing it; and (4) *Secondary supporting details*: add additional information or expand information given in primary supporting details. Students could earn up to 10 points, 1 point each for correctly recalling the topic and main idea and 4 points each for correctly providing the primary and secondary supporting details. Compared to the DIBELS 6th edition RTF measure, the RRR yields information regarding the quality and organization of information retold. This information can be useful in making instructional decisions.

Two preliminary studies have investigated the use the RRR for measuring reading comprehension of narrative (Shapiro, Fritschmann, Thomas, Hughes, & McDougal, 2010) and expository (Fritschmann, Shapiro, & Thomas, 2010) text. The initial investigation of the narrative version of the RRR was conducted with a different sample than the investigation of the expository version of the RRR, thus limiting direct comparisons of text type. The preliminary studies investigated the convergent validity, as well as the ability of the RRR to predict scores on the PSSA. An adapted version of the DIBELS 6th edition RTF measure was used, which combined the elements of the Fuchs et al. (1988) methods for calculating the total number of words retold and the Good and Kaminski (2002) and Roberts et al. (2005) method for scoring RTF. Specifically, participants were permitted to finish reading the entire passage before retelling the story for 1 minute. In addition, the passage remained in view during the retell, which is a deviation from the Fuchs et al., Good and Kaminski, and Roberts et al. studies.

At the third grade level, the highest correlations were between the RRR and Adapted RTF (Narrative winter r = .59 & spring r = .42, p < .01; Expository winter r = .55 & spring r = .46, p < .01). Weaker correlations were found between RRR with ORF (Narrative winter r = .23 & spring r = .21, p < .01; Expository winter r = .16 & spring r= .12, ns) and RRR with PSSA (Narrative winter r = .24 & spring r = .25, p < .01; Expository winter r = .02 & spring r = -.02, ns). Across the two studies, differences were noted in the magnitude and significance of correlations between RRR with ORF and PSSA. Differences were also noted in the results for the backwards elimination regression analysis for variables predicting third grade PSSA scores. For the narrative study, RRR added significantly (*p* < .05) to ORF's (*p* < .001) prediction of PSSA, with ORF and RRR accounting for 30% of the variance in explaining PSSA. Conversely, for the expository study, RRR did not add significantly to ORF's (*p* < .001) prediction of PSSA, with ORF alone accounting for 31% of the variance in explaining PSSA. It was speculated that having the passage present during the retell may have impacted the findings, with some participants more likely to copy directly from the text as opposed to engaging in more active and deeper processing (Hidi & Anderson, 1986). An additional limitation of these studies is the use of backwards elimination regression. Several problems have been identified with backwards elimination regression. In backwards elimination regression the order of elimination is based solely on the empirical relationship among the variables entered into the equation. As noted by Licht (1995) "pure empirical selection of predictors is likely to be highly sample specific and is not likely to include all theoretically relevant, or to exclude all irrelevant predictors. Thus, these procedures are likely to produce misleading and nonreproducible results" (p. 53).

Further research was warranted to examine the psychometric and diagnostic properties of the RRR measure in order to determine its usefulness as a screening measure of reading comprehension abilities. The reasons for this were (a) the need for valid, reliable, and sensitive measures for identifying children at risk for reading problems, (b) the limitations of standardized, norm-referenced reading tests for screening and instructional decision making, (c) the limitations of ORF in assessing reading comprehension, and (d) the limitations of existing retell measures including weak technical adequacy and instruction utility, and (e) the preliminary nature and mixed findings across the two previous investigations of the RRR. The purpose of this investigation was to evaluate the utility of using the RRR for identifying children at risk for reading comprehension difficulties with narrative and expository text. This study sought to replicate and broaden the scope of the previous investigations of the RRR through (a) allowing for direct comparison of text within the same study, (b) examining the convergent validity of the RRR with both CBM and standardized, norm-referenced measures, (c) examining the predictive validity of the RRR through use of logistic regression to investigate the degree to which the RRR was able to add to ORF's ability to accurately classify third grade students as proficient readers or non-proficient readers on the GRADE, 4Sight, and PSSA, (d) examining whether the RRR had a greater contribution than the Adapted RTF to ORF's prediction of students who had been identified as proficient readers or non-proficient readers on the GRADE, 4Sight, and PSSA, and (d) examining the alternate form reliability and interscorer reliability of ORF, RTF, and RRR.

Chapter Three: Methodology

Participants and Setting

This research study was approved by the Institutional Review Board at Lehigh University. A priori power analyses were conducted to determine the appropriate sample size for each statistical procedure by using Hsieh, Bloch, and Larsen's (1998) method for calculating sample size for logistic regression and Cohen's (1992) tables for calculation of the sample size for correlation. Results indicated that logistic regression of a binary dependent variable and two continuous independent variables required a sample size of 103 to achieve 80% power at a 0.05 significance level (Hintze, 2008; Hsieh et al.) and bivariate correlation using an alpha of .05, medium effect size, and power of .80, would require a sample size of 85 (Cohen).

The participants in this study were 107 elementary school children attending third grade in one public elementary school in Eastern Pennsylvania. All students in third grade (*n* = 127 students), including students with individualized education programs (IEP), were invited to participate in the study. A letter from the principal indicating her approval of the study and a consent form were sent home to each student's parent or guardian. Two rounds of consent forms were sent home via the student's classroom teacher, which yielded a 100% return rate. Sixteen parents/guardians did not give consent for their child to participate in the study. Student assent was obtained at the time of data collection. Four students did not give assent to participate in the study. The final sample consisted of 56 male and 51 female students ranging in age from 8 to 9 years old (*M* = 8 years 10 months old). The sample was predominantly Caucasian. The final sample included seven students with an IEP for a Specific Learning Disability (SLD) in Reading.

The elementary school's demographic characteristics can be found in Table 1. The school included students in grades three through six. Nineteen percent of students in the school had an IEP and forty-four percent of students in the school received free or reduced lunch. The school made adequate yearly progress with sixty-nine percent of students in the school performing in the proficient or advanced range on the Pennsylvania System of School Assessment (PSSA; Data Recognition Corporation [DRC], 2011) reading assessment.

Measures

Group Reading Assessment and Diagnostic Evaluation: Comprehension Composite (GRADE; Williams, 2001). The GRADE is a standardized, norm-referenced measure of reading ability for students in pre-kindergarten through adulthood (Williams). The GRADE was designed as a diagnostic tool to measure students' reading skills, chart progress, and monitor growth (Williams). At the third grade level, the GRADE assesses vocabulary (Word Reading and Vocabulary subtests) and comprehension (Sentence Comprehension and Passage Comprehension subtests), which together generate a Total Test score. The third grade level test also includes an Oral Language composite (Listening Comprehension subtest); however, this subtest is optional and does not contribute to the Total Test score. The GRADE is an untimed, group administered, multiple-choice test. The reported administration time for the third grade level total test is 1 to 2 hours with two test sessions recommended (Williams). For the purposes of this investigation, only the Comprehension composite (Sentence

Comprehension and Passage Comprehension subtests) was administered. The Sentence Comprehension subtest measures a student's ability to comprehend a sentence as a whole thought or unit (Williams). Students are required to silently read a single sentence with a missing word represented by a blank and choose one of four or five single-word choices to replace the blank (Williams). The Sentence Comprehension subtest draws on a student's knowledge of context clues, vocabulary, parts of speech, and sentence structure (Williams). The Passage Comprehension subtest measures a student's reading comprehension skills for a single paragraph or multiple paragraphs. Students are required to silently read a passage of one or more paragraphs and to answer three, four, or five questions about the passage each with four response choices (Williams). The Passage Comprehension subtest draws on a student's ability to apply the following metacognitive strategies: questioning, clarifying, summarizing, and predicting (Williams).

For this study, the raw score for the Comprehension composite of the GRADE was included as a criterion-measure for the convergent validity analysis of the Reading Retell Rubric (RRR). A Normal Curve Equivalent (NCE) of 40 was used to assign participants to the proficient reader (NCE ≥ 40) and non-proficient reader (NCE < 40) groups. This group assignment served as the categorical dependent variable for the diagnostic predictive validity analyses of the RRR. The cut scores for the GRADE were determined by Riedel (2007), who analyzed the diagnostic predictive validity of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS, 6th edition; Good & Kaminski, 2002) in predicting performance on the GRADE using both a NCE of 40 and performance at the 40th percentile. See Riedel for a complete description of the decision process for selecting a NCE of 40 as the criterion for group assignment.

The Comprehension composite of the GRADE was selected because it utilizes a method for assessing reading comprehension that is representative of summative assessments (e.g., statewide reading assessments and end-of-unit chapter tests) widely used in the United States (Torgesen et al., 2007) and because it incorporates both narrative and expository text. The publishers of the GRADE report good psychometric properties in the technical manual (Williams, 2001). The reported internal consistency estimates for the third grade level Sentence Comprehension subtest ranged from .83 to .87, with a reported split-half (i.e., odd/even) reliability ranging from .91 to .94 (Williams). Similarly, the reported internal consistency estimates for the third grade level Passage Comprehension subtest ranged from .83 to .85, with a reported split-half (i.e., odd/even) reliability ranging from .91 to .92 (Williams). The reported alternate form reliability for the third grade level total test was .94 and the test-retest reliability for a mean of 16.8 days was .93 (Williams). The reported convergent validity of the third grade level Total Test score with the Gates-MacGinitie Reading Tests (MacGinitie et al., 2000) was .86 and the predictive validity ranged from .76 to .77 for the second, fourth, and sixth grade levels of the GRADE with the TerraNova (CTB/McGraw-Hill, 2002) reading test (Williams).

Pennsylvania 4Sight Reading Benchmark Assessment (4Sight; Success for All Foundation, 2008). The 4Sight serves as a screening measure to predict students' performance on the PSSA reading assessment, a statewide assessment that is administered yearly to evaluate students' progress with grade level reading standards and serves as a measure of educational accountability (DRC, 2011). The 4Sight was designed to mimic the format of the PSSA, with respect to the standards represented on the state test, the weight/balance of the standards, the reporting scale used for the state test, the types of items, difficulty level of items, and types of distracter items (Success for All Foundation). Content validity was established through analysis of the blueprints for the PSSA reading assessment, assessment anchors, and released state assessments, practice items, and administration and scoring guides (Success for All Foundation).

The third grade 4Sight reading assessment is a one-hour, group administered test consisting of 29 multiple-choice items, each with four response choices, and 1 open-ended item. The multiple-choice items measure students' overall understanding of the passage, including setting, main idea, supporting details, sequence, and inferences (DRC, 2011). The open-ended item measures students' ability to prepare an answer, summarize information, and provide supporting details from the text in their response (DRC). At the third grade level, the 4Sight measures students' comprehension and reading skills (60-80% of test) and interpretation and analysis skills (20-40% of test) for narrative (50-70% of test) and expository (30-50% of test) text (DRC; Success for All Foundation, 2008).

There are 5 versions of the 4Sight that can be administered throughout the school year (fall, mid-fall, winter, late winter, and spring) leading up to the PSSA administration in the spring. The initial administration in fall represents the baseline score for the student. The elementary school participating in this study administered the benchmark test (fall), test 1 (mid-fall), and test 2 (winter) prior to the PSSA administration. Data from the winter administration were included in the analysis for this study because of its proximity to the administration of the PSSA. The total raw reading score from this assessment was included as a criterion-measure for the convergent validity analysis of the RRR. The predicted scaled score performance level cut score for the 4Sight was used to assign participants to the proficient reader and non-proficient reader groups. This group assignment served as the categorical dependent variable for the diagnostic predictive validity analyses of the RRR. The cut scores for the 4Sight were determined using those approved by the Pennsylvania State Board of Education (2007) for the PSSA third grade reading assessment, which are as follows: Advanced = 1442 or above, Proficient = 1235 to 1441, Basic = 1168 to 1234, and Below Basic = 1000 to 1167. Successful performance (i.e., Adequate Yearly Progress) is considered as proficient or advanced. See the "Pennsylvania System of School Assessment Grade 3 Reading Performance Level Descriptors" (Pennsylvania State Board of Education, 2005) for a complete description of the below basic, basic, proficient, and advanced performance levels.

The 4Sight was selected for this investigation because it is a well-established screening measure of students' reading performance on the PSSA. The psychometric characteristics (e.g., content and concurrent validity) of the 4Sight have been evaluated and showed the 4Sight to have strong psychometric properties consistent with the PSSA statewide reading assessments (Success for All Foundation, 2008). The current third grade 4Sight assessment was piloted in the spring of 2007 with all forms re-correlated with spring 2008 PSSA scores (Success for All Foundation). Concurrent validity across the 5 versions of the third grade 2008 4Sight and PSSA reading assessments ranged from .81 to .89 (Success for All Foundation). These correlations are based on samples ranging from 2,298 to 11,011 (Success for All Foundation). Comparison of the third grade 4Sight reading estimates to actual PSSA performance yielded similar percentages with 63% of students found to be in the proficient/advanced range on the 4Sight Reading assessment and 70% of students found to be in the proficient/advanced range on the PSSA (Success for All Foundation). Inter-form reliability for all 4Sight Reading tests across grades 3 to 11 ranged from .69 to .78, with the average inter-form correlation for the third grade reading test falling at .73 (average *n* = 46,400) (Success for All Foundation).

2011 Pennsylvania System of School Assessment Reading Assessment (PSSA; DRC, 2011). The PSSA is a statewide assessment that is administered yearly to all third grade students in the state where the study took place. The PSSA Reading assessment evaluates students' progress with grade level reading standards and serves as a measure of educational accountability in Pennsylvania (DRC). The third grade PSSA Reading assessment measures students' comprehension and reading skills (60–80% of test) and interpretation and analysis skills (20–40% of the test) for narrative (50–70% of the test) and expository (30–50% of the test) text (DRC). The PSSA Reading assessment includes both multiple-choice, each with four response choices, and open-ended questions. The multiple-choice items measure students overall understanding of the passage, including setting, main idea, supporting details, sequence, and inferences (DRC). The open-ended items measure students' ability to prepare an answer, summarize information, and provide supporting details from the text in their response (DRC).

The PSSA was selected for this investigation because it is the primary summative measure of students' reading performance (e.g., students' academic progress with state reading standards) within Pennsylvania. The total raw reading score from this assessment was included as a criterion-measure for the convergent validity analysis of the RRR. However, the school was unable to provide the raw scores for the PSSA. The PSSA scaled scores provided by the school were converted to raw scores using the *Raw-to-Scaled Score Conversion Table* provided in the Technical Report for the 2011 PSSA (DRC, 2011). The scaled score performance level cut score for the PSSA was used to assign participants to the proficient reader and non-proficient reader groups. This group assignment served as the categorical dependent variable for the diagnostic predictive validity analyses of the RRR. The cut scores for the PSSA were determined using those approved by the Pennsylvania State Board of Education (2007) for the PSSA third grade reading assessment, which are as follows: Advanced = 1442 or above, Proficient = 1235 to 1441, Basic = 1168 to 1234, and Below Basic = 1000 to 1167. Successful performance (i.e., Adequate Yearly Progress) is considered as proficient or advanced. See the "Pennsylvania System of School Assessment Grade 3 Reading Performance Level Descriptors" (Pennsylvania State Board of Education, 2005) for a complete description of the below basic, basic, proficient, and advanced performance levels. The psychometric characteristics (e.g., content validity, construct validity, item fit and calibration) of the PSSA have been extensively evaluated and showed the PSSA to have strong psychometric characteristics consistent with other statewide assessments (DRC).

Reading Passages. The reading passages in this study were drawn from a pool of six narrative passages and six expository passages used in the previous investigations of the RRR (Fritschmann, Shapiro, & Thomas, 2010; Shapiro, Fritschmann, Thomas, Hughes, & McDougal, 2010). All passages were originally written for the purposes of universal screening and progress monitoring of reading (i.e., ORF measurement) for students in third grade. The original narrative passages were selected from the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002) and

AIMSweb (Pearson Education Inc., 2008) passages where the key narrative story elements could be identified (e.g., characters, setting, plot, etc.), whereas the original expository passages were selected from edHelper (2009) passages where the key expository story elements could be identified (e.g., informational, main idea, supporting details, etc.).

A limitation identified in the Shapiro et al. study was the variability of passage difficulty. Although passages were selected from material commonly used for universal screening of reading, and the passages were carefully gauged to meet grade level readability requirements, there was more variability in the readability of passages than expected. To address this limitation, data from the Fritschmann et al. and Shapiro et al. studies were used to determine passage compatibility and readability. Specifically, the descriptive statistics, correlational analysis, and readability analysis (i.e., Spache Readability; Micro Power & Light Co., 2008; Lexile Analyzer; MetaMetrics Inc., 2008) were examined to carefully calibrate passage difficulty and identify compatible passages. Passage characteristics are presented in Table 2.

Oral Reading Fluency (ORF; Good & Kaminski, 2002; Shinn & Shinn, 2002). Each participant was required to read a passage aloud for one-minute. The number of words read correctly per one minute (WCPM) comprised the participant's performance score. This score was computed by subtracting any hesitations, mispronunciations, substitutions, omissions, and transpositions from the total number of words read in one minute. A reader was not penalized for insertions, repetitions, articulation and dialect, or self-corrections provided within three seconds. For this study, the median WCPM across three passages within each text type was included in the convergent and diagnostic predictive validity analyses. ORF was chosen because it is the most widely used and thoroughly investigated curriculum-based measure of reading ability. In addition, it is relatively easy to administer and score and can be used for both narrative and expository text.

Many studies have confirmed the technical adequacy of ORF (see Dynamic Measurement Group [DMG], 2008, 2011b and Shinn and Shinn, 2002, for a summary of the reliability and validity studies). In a comprehensive review of published and unpublished work examining the reliability and validity of ORF, Martson (1989) reported the following reliability and validity coefficients: test-retest reliability ranged from .82 to .97, parallel form reliability ranged from .84 to .89, criterion-related validity with published measure of reading competence ranged from .63 to .90, and interrater reliability was .99. The DMG (2011b) reported the following reliability and validity coefficients for ORF WCPM in the DIBELS Next technical manual: alternate-form reliability ranged from .92 to .98, test-retest reliability ranged from .91 to .97, inter-rater reliability was .99, and convergent validity with GRADE Total Test ranged from .61 to .75

Adapted Retell Fluency (Adapted RTF; Fritschmann et al., 2010; Good & Kaminski, 2002; Shapiro et al., 2010). The DIBELS 6th edition RTF measure was designed to be administered in combination with ORF to prevent children from focusing on fluency without attending to meaning and identify children whose comprehension is inconsistent with their ORF (DMG, 2011a; Good & Kaminski). The DIBELS 6th edition RTF measure was adapted for the purposes of this investigation. The participant is allotted one minute to retell the story. Whereas the original RTF measure is administered after the participant has read the passage for one minute, only assessing the information that the participant read in the one minute time period, the adapted version of the RTF was administered after the participant had finished reading the entire passage. This change is supported by Bellinger and DiPerna (2001), who indicated that the DIBELS 6th edition RTF measure “may be an insufficient measure of comprehension” because the assessment is based on a “short period of time” (i.e., 1-minute oral reading task), thus limiting the “amount of meaningful information that the child could comprehend” (p. 418).

Contrary to the preliminary investigations of the RRR, this investigation used a similar methodology to Fuchs et al. (1988), Good and Kaminski, and Roberts et al. by removing the passage from the participant’s view and then asking the participant to retell the story they just read in their own words. It is speculated that having the passage present during the retell may have impacted the findings of the preliminary investigations of the Adapted RTF and RRR, with some participants more likely to copy directly from the text as opposed to engaging in more active and deeper processing (Hidi & Anderson, 1986). In particular, Davey (1988) found that when the text was present struggling readers were more likely to use verbatim language in text recall, which could have potential mistakenly placed the individual in the low risk range for comprehension due to the high number of literal story facts or ideas retold. It is conjectured that these small changes in the methodology of the Adapted RTF may produce different outcomes from the previous investigations of the Adapted RTF.

In accordance with the DIBELS 6th edition RTF administration, the total number of words retold in 1 minute represented a participant’s Adapted RTF score. This score was computed by the examiner recording, as the participant was responding, the number of words that a reader could retell within a one minute time period. Points were awarded for words or sentences that were related to the topic. Mistakes or inconsistencies in the retell did not count as errors as long as the participant remained on topic. A participant did not earn points for exclamations (e.g., uhh, umm), songs or recitations, rote repetition, repeating ideas previously given in the retell, or irrelevant and off-track comments. For the DIBELS 6th edition RTF measure, a participant can earn up to 94 points. Given that the Adapted RTF measure was based on a larger sample of content (i.e., the entire passage instead of only the amount read in one-minute), the number of points a participant could earn was increased to up to 200 points (See Appendix B). For this study, the median total number of words retold in 1 minute across three passages within each text type was included in the convergent and diagnostic predictive validity analyses. Additionally, the Adapted RTF was audio recorded and transcribed to allow for comparison between the live Adapted RTF score and transcribed re-scoring of the total number of words retold in one minute.

Technical Adequacy information for the DIBELS 6th edition RTF measure was reported in the DIBELS Next Technical Manual (DMG, 2011b). Roberts et al. (2005) reported reliability of the RTF measure with students in first grade for retell of a single passage as $r = .57$ and for retell across multiple passages as $r = .87$. Four studies reported the concurrent criterion-related validity of the RTF measure with students in first grade: $r = .59$ and .68 with the Test of Word-Reading Efficiency (see Burke & Hagan-Burke, 2007), $r = .51$ with the GRADE (see Riedel, 2007), $r = .81$ and .42 with the Woodcock Diagnostic Reading Battery (see Roberts et al.), and with students in third grade: $r = .50$

with the Oregon State Assessment Test (see Mckenna & Good, 2003). Riedel reported the predictive criterion-related validity of the RTF measure with students in first grade as *r* = . 41 with the GRADE and *r* = .39 and .46 with the TerraNova.

Reading Retell Rubric (RRR; Fritschmann et al., 2010; Shapiro et al., 2010). The RRR was designed to be a CBM-like measure of reading comprehension. The RRR focuses on the key story structure elements that an individual includes in his or her oral retell. The RRR was administered in conjunction with the Adapted RTF measure. The information that the participant provided during the first minute of their retell, which was used to score the Adapted RTF measure, and any additional information provided beyond the 1 minute time frame was used to score the RRR. Since the Adapted RTF and RRR cannot be scored at the same time, for the purposes of this investigation, the examiner scored the Adapted RTF measure live and audio recorded the participant's retell to score RRR at a later time.

The narrative and expository RRR were developed through an extensive review of the literature (e.g., Best et al., 2008; Blachowicz & Ogle, 2008, 2001; Caldwell & Leslie, 2005; Cash & Schumm, 2006; Copmann & Griffith, 1994; Graesser et al., 2003; McKenna & Stahl, 2009; Medina & Pilonieta, 2006; Oakhill & Cain, 2007) and examination of narrative (DIBELS and AIMSweb) and expository (edHelper) reading passages to see what key story elements could be identified. A team of university professors and graduate research assistants independently reviewed the materials to identify a list of the most common narrative and expository story elements found in these sources. Similar items were grouped together into broad categories. Unique or single items were discussed with the whole research team to see if they should be made into their own category, collapsed into one of the existing categories, or excluded. Description of the narrative and expository RRR is as follows.

For the narrative RRR the student could earn up to 10 points for correctly providing each of the following story structure elements: (1) *Theme*: the central idea or point of the passage; (2) *Problem*: an obstacle or conflict the main character must resolve; (3) *Goal*: how the main character wants the problem to be resolved or what the main character is attempting to achieve; (4) *Setting*: where and when the story takes place; (5) *Characters:* people or animals in the story; (6) *Initiating event:* an idea or action that sets further events in motion or causes the main character to respond in some way; (7) *Climax*: when the conflict or problem is resolved; (8) *Sequence*: retelling the story in a structural or temporal order; (9) *Problem solution:* how the problem was resolved; and (10) *End of story:* conclusion or how the story turns out. For this study, the median total number of elements included in the retell across the three narrative passages was included in the convergent and diagnostic predictive validity analyses. Note each story structure element is worth 1 point (See Appendix C for an example of the narrative RRR record form).

For the expository RRR the student could earn up to 10 points for correctly providing each of the following story structure elements: (1) *Topic:* the subject of the text; (2) *Main idea*: what the text is all about, or what most of the sentences are about, or the overarching theme; (3) *Sequence*: retelling the story in a structural or logical order; (4) *Primary supporting details*: the facts needed to understand the main idea or support the main idea by explaining it and developing it; and (5) *Secondary supporting details*: adding additional information or expanding information given in primary supporting details. Topic, main idea, and sequence are each worth 1 point whereas primary supporting details are worth 4 points and secondary supporting details are worth 3 points. The original RRR for expository text did not include a category for sequence (e.g., students could earn 1 point each for topic and main idea and 4 points each for primary and secondary supporting details). After extensive review of the professional literature it was clear that the sequence in which a student retells the text provides important information about how he or she prioritizes, chunks, synthesizes, and expresses information (Blackowicz & Ogle, 2001). Therefore, sequence was added to the RRR for expository text as an additional category for this study. For this study, the median total number of elements included in the retell across the three expository passages was included in the convergent and diagnostic predictive validity analyses (See Appendix D for an example of the expository RRR record form).

The development of the RRR scoring template for each passage included several steps. First, narrative passages were extensively reviewed from both DIBELS 6th edition (Good & Kaminski, 2002) and AIMSweb (Pearson Education Inc., 2008) in order to identify passages in which the aforementioned key narrative story structure elements could be identified. Likewise, the expository passages from edHelper (2009) in which the key expository story structure elements listed above could be identified were comprehensively reviewed. Next, a team of doctoral level graduate research assistants independently read each passage and identified content from the passage to match each story structure element on the RRR. The graduate student reviews were compared to those created by the researchers. All responses were analyzed for consistency. Passages were eliminated if story structure elements were missing or consensus could not be reached. Once passages were selected, simulated retells were developed. Four graduate students scored the retells using the RRR. The reviewers were also instructed to write additional notes regarding areas of scoring difficulty or concerns. This information was utilized to refine the format and scoring of the RRR. Next, the narrative and expository RRR were piloted in two 3rd grade classrooms. Results of the pilot data collection were used to further refine and finalize the RRR templates for the 6 narrative and 6 expository passages used in the preliminary data collection (i.e., Fritschmann et al., 2010; Shapiro et al., 2010). See Appendix C and D for an example of a RRR scoring template for a narrative passage and an expository passage.

There is limited research on the psychometric properties of the RRR. Results from the preliminary investigation yield a mean interscorer agreement of 86% (range, 50 to 100%). At the third grade level, RRR for narrative and expository text correlated with ORF r = .12 to .26 and Adapted RTF r = .46 to .59 (Fritschmann et al., 2010; Shapiro et al., 2010).

Procedures

Recruitment. Participants were recruited by sending consent forms to parents/guardians through the student's classroom teacher. Consent forms were sent home one month prior to the testing session. A second round of consent forms were sent home two weeks later. Participants had the opportunity to provide their assent for participation in this study at the start of data collection.

Training. The principal investigator facilitated all training sessions. The third grade classroom teachers were trained to administer the Comprehension Composite of the GRADE using the administration manual provided by the publisher. The six classroom teachers received direct instruction and practice in administration during a group training session and an individual follow-up session. The school's intervention coordinator and a graduate research assistant were trained to score the Comprehension Composite of the GRADE using the scoring manual provided by the publisher. The intervention coordinator and a graduate research assistant each received direct instruction and practice in scoring during an individual training session. Five graduate research assistants (i.e., data collectors) were trained to administer and score the ORF, Adapted RTF, and RRR. Data collectors were provided with a set of standardized directions for administration and scoring. All data collectors completed online video training through the *DIBELS Training Institute: Essential Workshop* (DMG, 2007) module 8 for administration and scoring of DIBELS ORF and module 9 for administration and scoring of RTF. Each data collector also received direct instruction and practice using audio recordings of students' oral reading and retell. Each data collector was required to achieve a criterion score within two words on ORF, two words on Adapted RTF, and two points on RRR for the three narrative and three expository passages prior to data collection. The school was responsible for training individuals to administer and score the 4Sight and PSSA assessments.

Testing sessions. Testing sessions were conducted at the end of January and beginning of February immediately preceding the winter 4Sight administration in February and PSSA administration in March. Each participant took part in five testing sessions. First, each third grade classroom teacher administered the Comprehension composite of the GRADE to his or her class. A graduate research assistant scored the GRADE student booklets for students who participated in this study. The school's intervention coordinator scored the GRADE student booklets for students who did not have consent to participate in this study. Next, graduate research assistants administered the ORF, Adapted RTF, and RRR to the participants over two sessions. The order of the type of text (i.e., narrative vs. expository) and passage within each text type (i.e., narrative passage 1, 2, and 3 or expository passage 1, 2, and 3) was counterbalanced. The graduate research assistants scored the ORF and RTF during the live administration. The participant's oral retell for each passage was audio recorded to allow the graduate research assistants to score the RRR at a later time. In addition, following the procedures of Pressley, Hilden, and Shankland (2005) the Adapted RTF recordings were transcribed to allow for comparison between the live Adapted RTF score and transcribed re-scoring of the total number of words retold. Fourth, each classroom teacher administered and scored the 4Sight reading assessment. During the final testing session, each classroom teacher administered the PSSA. The school's intervention coordinator provided the principal investigator with the 4Sight and PSSA scores for the students who participated in this study.

Testing procedures. The classroom teachers followed the standardized administration procedures for the Comprehension composite of the GRADE provided by the publisher in the administration manual. The classroom teacher filled out the identifying information on the front cover of the student booklets. The student booklets and pencils were distributed to the class. The classroom teacher read the standardized directions in the administration manual to the students. The classroom teacher administered the sentence comprehension subtest, followed by the passage comprehension subtest. During testing, the teacher checked to make sure students were marking their booklets in the correct manner (e.g., circling one response choice for each item). Upon completion, the classroom teacher collected the student booklets and delivered them to the school's intervention coordinator. The school's intervention coordinator separated out the booklets for the students who had consent to participate in this study and provided these to the principle investigator of this study. A graduate research assistant scored the GRADE student booklets for the students who had consent to participate in this study. The school's intervention coordinator scored the student booklets for students who did not have consent to participate in this study.

Next, the order of the type of text (i.e., narrative vs. expository) and passage within each text type (i.e., narrative passage 1, 2, and 3 or expository passage 1, 2, and 3) was counterbalanced. Participants were randomly assigned to 1 of 6 narrative record forms and 1 of 6 expository record forms. Administration procedures for ORF and Adapted RTF were modified from DIBELS 6th edition (Good & Kaminski, 2002) and AIMSweb (Shinn & Shinn, 2002). The examiner placed the record from on a clipboard and positioned it so that the participant could not see what the examiner recorded. The examiner placed a copy of the passage in front of the participant. The examiner said these directions verbatim: *"Please read this (pointed to passage in front of participant) out loud. If you come to a word you don't know, I will tell you the word so you can keep reading. When I say stop, I will ask you to tell me about what you read, so do your best reading. Start here (pointed to the first word of the passage). Begin."* The examiner started the timer when the participant said the first word of the passage (not the title). If the participant failed to say the first word of the passage after three seconds, the examiner told the participant the word, marked it as incorrect, and then started the timer. The examiner timed for one minute. During the one minute time period, the examiner recorded errors by putting a slash (/) over each word read incorrectly. Errors included hesitations, mispronunciations, substitutions, omissions, and transpositions. If the participant hesitated or struggled with a word for three seconds, the examiner told the participant the word and marked it as incorrect. If the participant self-corrected within three seconds, the examiner did not count the word as an error. The examiner did not count insertions or repetitions as errors. Participants did not lose points for articulation or dialect (e.g., consistently pronounced “s” as “th,” said “retht” for “rest”). At the end of one minute, the examiner placed a bracket (]) after the last word provided by the participant. To score ORF, the examiner counted the total number of words read correctly in one minute. The formula for computing this score was: total number of words read in one minute minus errors made in one minute equals the total number of words read correctly in one minute.

At the end of one minute, the examiner prompted the participant to finish reading the passage by saying verbatim: *“Keep Reading.”* When the participant finished reading the passage, the examiner removed the passage from view. The examiner said these directions verbatim: *“Please tell me about what you just read in your own words. Try to tell me everything you can remember about the story. Ready, Begin.”* The examiner started the timer after he or she said *“Begin.”* The examiner timed for one minute. The examiner counted the number of words the participant retold by moving his or her pencil through the numbers as the participant was responding. Mistakes or inconsistencies in the retell did not count as errors as long as the participant remained on topic. The examiner did not score hesitations (e.g., umm, ah, like), songs or recitations, rote repetition, repeating ideas previously given in retell, or irrelevant and off-track information. The examiner prompted the participant one time within the one minute time period the first time the participant did not say anything for three seconds, by saying verbatim: *"Try to tell me everything you can remember about the story."* After this prompt, if the participant did not say anything or got off track for five seconds, the examiner circled the number of words in the participant's retell and discontinued the Adapted RTF task. Alternatively, if the participant reached the end of one minute, the examiner put a circle around the total number of words the participant retold.

Regardless of whether the participant stopped early because of the five second rule or ended at one minute, after completion of the Adapted RTF task, the examiner prompted the participant to finish retelling or add to their retell by saying verbatim: *"Keep going," "Try to tell me everything you can remember about the story,"* or *"Is there anything more you can tell me about the story."* The examiner used the information that the participant provided during and after the Adapted RTF to score the RRR. The examiner scored the RRR post-administration. The examiner could listen to the audio recording up to two times while using the scoring template to record the participant's retell on the RRR record form. For the Narrative RRR, the examiner circled zero if the participant omitted the item and one if the participant included the item in his or her retell. Each of the following ten items were worth one point: theme, problem, goal, setting, characters, initiating events, climax, sequence, problem solution, and end of the story. Partial responses were scored a one, such that every character or every aspect of the setting did not need to be included in order to receive one point. The Expository RRR consisted of five items: topic, main idea, primary supporting details, secondary supporting details, and sequence. For topic, main idea, and sequence the examiner circled zero if the participant omitted the item or part of the item in his/her retell and circled one if the participant stated the item in his or her retell. For primary supporting details, the examiner circled zero if the participant omitted the item in his or her retell, circled one of the participant provided one detail, circled two if the participant provided two details, circled three if the participant provided three details, or circled four if the participant provided four or more details. Similarly, for secondary supporting details, the examiner circled zero if the participant omitted the item in his or her retell, circled one of the participant provided one detail, circled two if the participant provided two details, or circled three if the participant provided three or more details. For both the Narrative and Expository RRR, the examiner calculated the total score by adding all of the points earned. In addition, an independent data collector transcribed and re-scored the Adapted RTF to allow for comparison between the live Adapted RTF score and transcribed re-scoring of the total number of words retold.

Procedural integrity. Procedural integrity was checked for administration of the GRADE using a self-assessment checklist. Teachers were asked to initial whether they had all of the necessary materials, instructions were given verbatim, and subtests were administered in accordance with the standard directions and procedures outlined by the GRADE manual. Space was provided for teachers to indicate any deviations from the standard administration.

Procedural integrity was also checked two times for each data collector. An observation checklist was used to check whether the examiner had all the necessary materials, instructions were given verbatim, and measures were properly administered and accurately scored. An independent observer noted whether procedures were followed on a step-by-step basis. If a lack of integrity was evident (the examiner earned less than 90% on the procedural integrity checklist), the examiner received additional training and guidance prior to collecting further data. An additional procedural integrity check was conducted to ensure the accuracy of the data collection.

Interscorer agreement and data entry checks. Interscorer agreement was assessed for 100% of the cases. An independent examiner separately scored the GRADE, ORF, RTF, and RRR. The independent examiner listened to audio recordings of the live administration of ORF, RTF, and RRR. For RTF, the independent examiner transcribed the retell and counted the total number of words retold. Interscorer agreement was first determined on a point-by-point basis. The total percentage of agreement was calculated for each measure (e.g., number of agreements divided by number of agreements plus disagreements multiplied by 100). Scoring discrepancies were resolved via a third independent examiner. Interscorer agreement for ORF, RTF, and RRR was also determined by calculating the percentage of interscorer checks that were within the 2-point criterion for judging interscorer agreement set by Good and Kaminski (2002). Data entry was also checked for 100% of the cases. An independent examiner checked to make sure data were entered accurately. Any data entry errors were noted and corrected.

Data Analyses

Preliminary analyses. The data were screened to check for violations of the assumptions underlying Pearson Product-Moment Correlations and Hierarchical Binary Logistic Regression. For Pearson-Product Moment Correlations, data were screened for (a) missing data, (b) univariate outliers, (c) normality, (d) linearity, and (e) homoscedasticity, using descriptive statistics, frequency analysis, missing value analysis (MVA), histograms, scatterplots, and Means comparison. For Hierarchical Binary Logistic Regression, data were screened for (a) missing data, (b) univariate outliers, (c) multivariate outliers, and (d) multicollinearity, using descriptive statistics, frequency analysis, MVA, Mahalanobis distance, and correlations.

Prior to data analysis, the best approach for dealing with missing data, violations or normality, and outliers was determined. It was decided that missing data would be addressed as follows: (a) if data were determined to be missing completely at random and removal of the missing cases would not impact statistical power (Schlomer, Bauman, & Card, 2010), then listwise deletion would be considered or (b) if statistical power was in jeopardy, multiple imputation or full information maximum likelihood methods would be considered (see Baraldi & Enders, 2010; Schlomer et al.). It was also decided that violation of normality, in particular the skewness of the data, would be evaluated for severity, and if necessary, statistical transformation would be considered. Finally, it was decided that if outliers emerged, the data would be checked to ensure the outlier score was genuine, not just a data entry error, and if a genuine outlier was identified, then statistical literature (e.g., Tabachnick & Fidell, 2007) would be consulted to determine the best course of action (e.g., remove outliers from data file or recoding the value).

Pearson product-moment correlations. Correlational analyses were conducted to examine the first and second research questions which were as follows: (a) what is the convergent validity of the RRR for assessing reading comprehension of narrative text with other measures typically used to assess narrative reading comprehension and (b) what is the convergent validity of the RRR for assessing reading comprehension of expository text with other measures typically used to assess expository reading comprehension. Investigations of convergent validity examine the extent to which two measures assess similar constructs. Consequently, the validity of the RRR was measured by the extent to which the RRR correlated with other measure of reading ability (i.e., ORF, Adapted RTF, GRADE, 4Sight, and PSSA).

Group assignment. In order to evaluate the predictive validity of the RRR, ORF, and Adapted RTF to the GRADE, 4Sight, and PSSA participants were assigned to two groups: (a) proficient readers and (b) non-proficient readers. This group assignment served as the categorical dependent variable for the diagnostic predictive validity analyses (i.e., proficient readers = 0 and non-proficient readers = 1). For the GRADE, group assignment was based on a NCE of 40. This cut score for the GRADE was determined by Riedel (2007). Students in the proficient reader group had a NCE greater than or equal to 40. Students in the non-proficient reader group had a NCE less than 40. For the 4Sight, group assignment was based on each participant's performance on the 4Sight reading assessment administered in winter. For the PSSA, group assignment was based on each participant's performance on the PSSA reading assessment administered in spring. The cut scores for the 4Sight and PSSA, which were determined using those approved by the Pennsylvania State Board of Education (2007) for the PSSA third grade reading assessment, were used to determine group assignment. Students in the proficient reader group (i.e., proficient and advanced readers) had a scaled score in the range of 1235 to 1442 or above on the 4Sight and PSSA. Students in the non-proficient reader group (i.e., basic and below basic readers) had a scaled score between 1000 and 1234 on the 4Sight and PSSA.

Hierarchical binary logistic regression. Regression analysis was conducted to examine the third, fourth, fifth, and sixth research questions, which are as follows: (a) does the RRR for assessing reading comprehension of narrative text improve ORF's prediction of students who are proficient readers and those who have been identified as non-proficient readers on the GRADE, 4Sight, or PSSA, (b) does the RRR for assessing reading comprehension of expository text improve ORF's prediction of students who are proficient readers and those who have been identified as non-proficient readers on the GRADE, 4Sight, or PSSA, (c) does the RRR for assessing reading comprehension of narrative text have a greater contribution to ORF's prediction of students who are proficient readers and those who have been identified as non-proficient readers on the GRADE, 4Sight, or PSSA, as compared to Adapted RTF, and (d) does the RRR for assessing reading comprehension of expository text have a greater contribution to ORF's prediction of students who are proficient readers and those who have been identified as non-proficient readers on the GRADE, 4Sight, or PSSA, as compared to Adapted RTF. Hierarchical Binary Logistic Regression was used to measure how well ORF, RRR, and Adapted RTF predict performance (i.e., proficient versus non-proficient readers) on the GRADE, 4Sight, or PSSA. The analyses examined the ability of ORF to predict performance on the GRADE, 4Sight, or PSSA alone, as well as the additive impact of RRR or Adapted RTF on ORF's ability to predict performance on the GRADE, 4Sight, or PSSA. The level of measurement included a dichotomous categorical dependent variable (i.e., proficient readers = 0 and non-proficient readers = 1) and two continuous predictor variables. The enter method was used, with all predictor variables being tested in two blocks (e.g., ORF alone and ORF combined with RRR or RTF). A Bonferroni correction was used to counteract the use of multiple comparisons for each research question (*p* < .017). The inverse of the odds ratio was calculated to indicate for every one unit increase in the independent variable score, how many times less likely participants were to be categorized as non-proficient readers on the dependent variable.

The hierarchical binary logistic regression analysis examined sensitivity, specificity, positive predictive power, negative predictive power, and odds ratio. Sensitivity refers to the number of true positives or the percentage of participants accurately identified by the model as non-proficient readers on the GRADE, 4Sight, or PSSA. Specificity refers to the number of true negatives or the percentage of participants accurately identified by the model as proficient readers on the GRADE, 4Sight, or PSSA. Positive predictive power refers to the percentage of cases that the model predicted to be non-proficient readers on the GRADE, 4Sight, or PSSA and that were actually non-proficient readers on the GRADE, 4Sight, or PSSA. Negative predictive power refers to the percentage of cases the model predicted to be proficient readers on the GRADE, 4Sight, or PSSA and that were actually proficient readers on the GRADE, 4Sight, or PSSA. The odds ratio refers to the probability of success (i.e., identified as proficient reader) over the probability of failure (i.e., identified as non-proficient reader).

Reliability. Reliability analysis was conducted to examine the alternate form reliability of ORF, Adapted RTF, and RRR and the interscorer agreement for the GRADE, ORF, Adapted RTF, and RRR. Correlations between the three narrative forms and between the three expository forms of the ORF, Adapted RTF, and RRR were examined respectively to determine the alternate form reliability of the ORF, Adapted RTF, and RRR. Interscorer agreement was first determined for the GRADE, ORF, Adapted RTF, and RRR on a point-by-point basis, yielding a total percentage of agreement score (i.e., number of agreements divided by number of agreements plus disagreements multiplied by 100). Interscorer agreement was also determined for ORF, Adapted RTF, and RRR using criteria outlined by the Good and Kaminski (2002), which indicates that both assessors should be within 2 points of each other on the final score.

Chapter Four: Results

The purpose of this investigation was to further examine the utility of using the RRR for identifying children at risk for reading comprehension difficulties with narrative and expository text. Specifically, this study investigated the convergent validity of the RRR by comparing performance on the RRR with performance on other established measures of reading comprehension administered at the same point in time. In addition, this study examined the ability of the RRR to enhance ORF's identification of children at risk for reading comprehension difficulties with narrative and expository text.

Data Screening

The data were screened to check for violations of the assumptions underlying Pearson Product-Moment Correlation (i.e., missing data, univariate outliers, normality, linearity, and homoscedasticity) and Hierarchical Binary Logistic Regression (i.e., missing data, univariate outliers, multivariate outliers, and multicollinearity). Descriptive statistics for all variables are presented in Table 3.

Missing data. The data were first screened for missing data. There was a complete data set for ORF, Adapted RTF, RRR, and 4Sight. The GRADE data was missing from two participants and the PSSA data was missing from one participant. Missing Value Analysis (MVA) was conducted to assess the extent and nature of missing data for each variable. There were no variables with 5% or more missing values; consequently, removal of the missing cases would not impact statistical power; therefore, missing data were dealt with using pairwise deletion for correlations and listwise deletion for logistic regression (Schlomer, Bauman, & Card, 2010; Tabachnick & Fidell, 2007).

Univariate outliers. Histograms, scatter plots, and descriptive statistics were examined for each variable to identify unviariate outliers. Univariate outliers were also assessed by transforming raw scores to z scores for all study variables. The z scores were then compared to a critical value of +/- 3.29 (*p* < .001; Tabachnik & Fidell, 2007). Scores that exceeded this critical value were over three standard deviations above or below the mean. This would indicate that the score was extreme and that it should be further evaluated to determine if it was part of the population. One univariate outlier (z score = 4.06) was identified for the Adapted RTF Expository variable. The highest score on Adapted RTF Expository was 139 words retold within one minute, which was 4 standard deviations above the mean; the next highest score was 101 words retold within one minute. The score of 139 words retold within one minute was also much higher than the score for the same participant on Adapted RTF Narrative, which was 108 words retold within one minute. After investigation, it was determined that the case was not part of the population and thus it was removed. After removal of the univariate outlier the sample size used for the correlation analysis for ORF, Adapted RTF, and RRR for both narrative and expository text and 4Sight was 106, 105 for PSSA, and 104 for GRADE. The sample sizes used for the logistic regression analyses with ORF, Adapted RTF, and RRR for narrative text were 105 with GRADE, 107 with 4sight, and 106 with PSSA. The sample sizes used for the logistic regression analyses with ORF, Adapted RTF, and RRR for expository text were 104 with GRADE, 106 with 4Sight, and 105 with PSSA.

Multivariate outliers. After removing the univariate outliers, the data were screened for multivariate outliers using Mahalanobis distance (Tabachnik & Fidell, 2007). Mahalnobis distance scores were requested via multiple regression analysis for each hierarchical binary logistic regression hypothesis. The Mahalanobis distance values were then compared to a critical value of χ^2 = 13.816 (p < .001; Tabachnik & Fidell). No multivariate outliers were detected.

Normality. Normality was assessed by transforming raw scores to z scores for all study variables. Normality of the z scores was first checked via visual inspection of histograms with imposed normal curves, and all but one variable appeared to be normally distributed (i.e., GRADE Comprehension Composite Raw Score). In addition, the skewness and kurtosis of the z scores was checked by dividing each variable's skewness and kurtoisis statistic by their respective standard error. The result was then compared to a critical value of +/- 3.29 (p < .001; Tabachnik & Fidell, 2007). Any z skewness or z kurtosis coefficients that exceeded this critical value were considered non-normal (Tabachnick & Fidell). All the z skewness and z kurtosis coefficients were below this critical value with one exception being the GRADE, where the distribution was found to be negatively skewed. In other words, there were relatively few low scores on the GRADE. Several transformations of the GRADE were attempted, including square root, logarithm, and inverse. Visual inspection of the histogram with imposed normal curve and z skewness and z kurtosis coefficients indicated that with the square root transformation the GRADE was normally distributed. Correlational analyses were conducted with both the GRADE raw score and the GRADE transformed via square root score.

Linearity. The linear relationship among the variables was examined using a scatterplot matrix. Examination of the scatterplots indicated that there were no curvilinear relationships (Tabachnick & Fidell, 2007). The linear relationship between variables was further investigated using SPSS Means. If the significant value for the Deviations from Linearity statistic was less than 0.05, then the relationship between the two variables was not linear. Deviations from linearity were identified between the GRADE Comprehension Composite Raw Score with RRR Narrative (*p* = .044), Adapted RTF Expository (*p* = .010), 4Sight (*p* = .006), and PSSA (*p* <.001). After square root transformation of the GRADE, the linearity among the variables improved; the deviation from linearity statistic was greater than 0.05 between the GRADE transformed via square root with RRR Narrative (*p* = .692), 4Sight (*p* = .106), and PSSA (*p* = .146). Transformation of the GRADE did not improve its linear relationship with Adapted RTF Expository (Deviation from Linearity *p* = .024). A deviation from linearity was also noted between ORF Expository and PSSA (*p* = .013). Deviation from linearity, specifically curvilinear relationships, is problematic because Pearson's *r* only captures linear relationships (Tabachnick & Fidell). Although the relationships between (a) the GRADE and Adapted RTF Expository and (b) ORF Expository and PSSA deviated from linearity, the relationships were *not* curvilinear; therefore, the variables were retained for the analysis.

Homoscedasticity. The scatterplot matrix was also used to check for homoscedasticity. Some of the bivariate scatterplots appeared slightly heteroscedastic. According to Tabachnick & Fidell (2007), "heteroscedasticity is not fatal to an analysis" (p. 85). Transformations of the variables were not conducted to improve homoscedasticity because the loss of interpretability did not seem worthwhile given that the relationships were only slightly heteroscedastic.

Multicollinearity. Multicollinearity was assessed by examining the correlations among the predictor variables (see Table 4). Correlations between all predictor variables were low (range, *r* = .23 to .36, p < .05; Evans, 1996). All correlations were less than the critical value of *r* = .80, thus, ruling-out multicollinearity (Grimm & Yarnold, 1995). Since some predictors were moderately correlated, the tolerance level and the Variance Inflation Factor (VIF) were also examined to rule-out mulitcollinearity. All tolerance values were greater than the critical value of .10 (range, .87 to .95) and all VIF values were less than the critical value of 10 (range, 1.06 to 1.15); therefore, multicollinearity did not appear to be of concern (Pallant, 2007).

Reliability Analyses

Alternate Form Reliability was assessed by examining correlations between the three narrative forms and between the three expository forms of the ORF, Adapted RTF, and RRR respectively. For the Narrative passages (i.e., The New Sofa, The Magic Fish, and The Surprise Party), the average alternate-form reliability of ORF was *r* = .92 (range, .91 to .94; *p* <.01), Adapted RTF was *r* = .67 (range, .66 to .67; *p* <.01), and RRR was *r* = .57 (range, .52 to .62; *p* <.01). For the Expository passages (i.e., Giraffes, Flamingos, and Owls), the average alternate-form reliability of ORF was *r* = .92 (range, .91 to .94; *p* <.01), Adapted RTF was *r* = .63 (range, .59 to .66; *p* <.01), and RRR was *r* = .52 (range, .41 to .58; *p* <.01).

Interscorer agreement was first determined for the GRADE, ORF, Adapted RTF, and RRR on a point-by-point basis, yielding a total percentage of agreement score. The total percentage of agreement for the GRADE was 100%. For the Narrative passages, the average percentage of agreement score for ORF was 99% (range, 86–100%), Adapted

RTF was 93% (range, 39–100%), and RRR was 90% (range, 50–100%). For the Expository passages, the average percentage of agreement score for ORF was 99% (range, 88–100%), Adapted RTF was 92% (range, 29–100%), and RRR was 91% (range, 50–100%).

Interscorer agreement was also determined for ORF, Adapted RTF, and RRR using the 2-point criterion for judging interscorer agreement set by Good and Kaminski (2002), which indicates that both assessors should be within 2 points of each other on the final score. For the Narrative passage, the average percentage of interscorer checks that were within 2-points of one another was 89% for ORF (range, 0 to 7 words), 43% for Adapted RTF (range, 0 to 35 words), and 93% for RRR (range, 0 to 5 points). For the Expository passage, the average percentage of interscorer checks that were within 2-points of one another was 93% for ORF (range, 0 to 7 words), 56% for Adapted RTF (range, 0 to 37 words), and 91% for RRR (range, 0 to 5 points).

Direct Comparison of Text Type

Direct comparison of text type yielded higher mean scores for the narrative passages across ORF, Adapted RTF, and RRR (See Table 3). The mean difference was 3.21 for ORF, 20.43 for Adapted RTF, and 0.50 for RRR. Paired-sample t-tests were conducted to evaluate the mean difference between the narrative and expository text versions of ORF, Adapted RTF, and RRR. There was a statistically significant difference between the narrative and expository text versions of ORF (*p* = .003), Adapted RTF (*p* < .001), and RRR (*p* = .008).

Convergent Validity Analyses

Research questions 1 and 2. Correlational analysis was used to answer the first and second research questions, which sought to investigate the convergent validity of the RRR by comparing performance on the RRR with performance on other established measures of reading comprehension administered at the same point in time. The median score across the three narrative and three expository passages was used for the correlations involving ORF, Adapted RTF, and RRR, whereas the raw score was used for correlations involving the GRADE, 4Sight, and PSSA. Due to violations to the assumptions of normality and linearity, the GRADE was transformed via square root. The correlations between the other measures with the GRADE raw score and the GRADE transformed via square root were nearly identical. Given that transformation limits interpretability (Tabachnick & Fidell, 2007) the raw scores of the GRADE were retained for the analyses. Correlations between the variables are presented in Table 4. All correlations were statistically significant (p < .05 or p < .01) and all variables were positively correlated. The strongest relationships (range, r = .72 to .94) were observed between ORF Narrative, ORF Expository, GRADE, 4Sight, and PSSA. Strong relationships were also observed between Adapted RTF Narrative and Expository (r = .62), Adapted RTF Narrative and RRR Narrative (r = .62), and Adapted RTF Expository with RRR Expository (r = .63).

Predictive Validity Analyses

Twelve hierarchical binary logistic regression analyses were conducted to determine which measures were significant predictors of each criterion variable (See Table 6 for a summary of findings across the twelve hierarchical binary logistic regression analyses). A Bonferroni correction was used to counteract the use of multiple comparisons for each research question ($p < .017$). For all of the analyses, ORF was entered into block 1 because of its documented strength as a predictor of performance on summative reading assessments, and RRR or Adapted RTF were entered into block 2 to examine their additive benefit. RRR and Adapted RTF were tested in parallel analyses to compare their unique contribution to ORF's ability to predict proficient and non-proficient readers on the criterion variable. These analyses were conducted for two purposes: (1) to determine whether RRR significantly added to ORF's ability to predict performance on each criterion variable and (2) to determine whether RRR had a greater contribution to ORF's ability to predict performance on each criterion variable as compared to Adapted RTF.

Results for the hierarchical binary logistic regression analyses are displayed in Tables 6 through 30. In each analysis, ORF was a statistically significant predictor of performance on the criterion variable. ORF's overall (a) sensitivity (i.e., percentage of participants accurately identified as non-proficient readers) ranged from 77% to 87% across the three criterion variables, (b) specificity (i.e., percentage of participants accurately identified as proficient readers) ranged from 87% to 91% across the three criterion variables, (c) positive predictive power (i.e., percentage of cases the model predicted to be non-proficient readers and were actually non-proficient readers) ranged from 65% to 70% across the three criterion variables, and (d) negative predictive power (i.e., percentage of cases the model predicted to be proficient readers and were actually proficient readers) ranged from 92% to 96%. For two of the analyses involving prediction of the GRADE, RRR Narrative and Adapted RTF Narrative produced a significant increase in ORF Narrative's predictive accuracy (increased predictive accuracy by 4.7%). For one analysis involving prediction of the PSSA, RRR Narrative also produced a significant increase in ORF Narrative's predictive accuracy (increase predictive accuracy by 1.9%).

Research question 3A. The first analysis assessed the predictive values of ORF Narrative and RRR Narrative with GRADE (See Tables 7 and 8). After controlling for ORF Narrative scores, the full model containing RRR Narrative was statistically significant, χ^2 (1, n = 105) = 14.867, $p < .001$, indicating that the model was able to distinguish between students who scored in the proficient range and students who scored in the non-proficient range on the GRADE. This model as a whole explained between 52.2% (Cox and Snell R Square) and 74.8% (Nagelkerke R Square) of the variance in reading status, and correctly classified 89.5% of the cases. Compared to model one which explained between 44.9% and 64.4% of the variance in reading status, correctly classifying 84.8% of the cases; RRR Narrative increased classification accuracy by 4.7%. Specifically, the addition of RRR improved (a) sensitivity (Model 1 = 77%; Model 2 = 88%), (b) specificity (Model 1 = 87%; Model 2 = 90%), (c) positive predictive power (Model 1 = 67%; Model 2 = 73%), and (d) negative predictive power (Model 1 = 92%; Model 2 = 96%). The odds ratio indicates that for every one unit increase in RRR Narrative scores, participants were 2.17 times less likely to be categorized as non-proficient readers on the GRADE. Similarly, for every one unit increase in ORF Narrative scores, participants were 1.10 times less likely be categorized as non-proficient readers on the GRADE.

Research question 3B. The second analysis assessed the predictive values of ORF Narrative and RRR Narrative with 4Sight (See Tables 9 and 10). The full model was not significant, χ^2 (1, *n* = 107) = 2.421, *p* = .120, indicating that RRR Narrative did not significantly add to ORF Narrative's ability to distinguish between students who scored in the proficient range and students who scored in the non-proficient range on the 4Sight (increased predictive accuracy by 1%). The model containing only ORF Narrative was statistically significant, χ^2 (1, *n* = 107) = 41.808, *p* < .001, explaining between 32.3% (Cox and Snell R Square) and 48.3% (Nagelkerke R Square) of the variance in reading status, and correctly classified 89.7% of the cases. The odds ratio indicates that for every one unit increase in ORF Narrative scores, participants were 1.06 times less likely to be categorized as non-proficient readers on the 4Sight.

Research question 3C. The third analysis assessed the predictive values of ORF Narrative and RRR Narrative with PSSA (See Tables 11 and 12). After controlling for ORF Narrative scores, the full model containing RRR Narrative was statistically significant, χ^2 (1, *n* = 106) = 5.393, *p* < .05, indicating that the model was able to distinguish between students who scored in the proficient range and students who scored in the non-proficient range on the PSSA. This model as a whole explained between 43.2% (Cox and Snell R Square) and 63.6% (Nagelkerke R Square) of the variance in reading status, and correctly classified 91.5% of the cases. Compared to model one which explained between 40.2% and 59.3% of the variance in reading status, correctly classifying 89.6% of the cases, RRR Narrative increased classification accuracy by 1.9%. Specifically, the addition of RRR improved (a) sensitivity (Model 1 = 86%; Model 2 = 91%), (b) specificity (Model 1 = 90%; Model 2 = 92%), (c) positive predictive power

(Model 1 = 70%; Model 2 = 74%), and (d) negative predictive power (Model 1 = 96%; Model 2 = 97%). However, using the Bonferroni corrected alpha level ($p < .017$), RRR was not identified as a significant predictor ($p = .025$). The odds ratio indicates that for every one unit increase in RRR Narrative scores, participants were 1.52 times less likely to be categorized as non-proficient readers on the PSSA. Similarly, for every one unit increase in ORF Narrative scores, participants were 1.08 times less likely to be categorized as non-proficient readers on the PSSA.

Research question 4A. The fourth analysis assessed the predictive values of ORF Expository and RRR Expository with GRADE (See Tables 13 and 14). The full model was not significant, χ^2 (1, n = 104) = 2.865, p = .091, indicating that RRR Expository did not significantly add to ORF Expository's ability to distinguish between students who scored in the proficient range and students who scored in the non-proficient range on the GRADE (no change in predictive accuracy). The model containing only ORF Expository was statistically significant, χ^2 (1, n = 104) = 57.821, $p < .001$, explaining between 42.6% (Cox and Snell R Square) and 61.0% (Nagelkerke R Square) of the variance in reading status, and correctly classified 87.5% of the cases. The odds ratio indicates that for every one unit increase in ORF Expository scores, participants were 1.08 times less likely to be categorized as non-proficient readers on the GRADE.

Research question 4B. The fifth analysis assessed the predictive values of ORF Expository and RRR Expository with 4Sight (See Tables 15 and 16). The full model was not significant, χ^2 (1, n = 106) = 3.316, p = .069, indicating that RRR Expository did not significantly add to ORF Expository's ability to distinguish between students who scored in the proficient range and students who scored in the non-proficient range on the 4Sight (increased predictive accuracy by 1%). The model containing only ORF Expository was statistically significant, χ^2 (1, *n* = 106) = 42.088, *p* < .001, explaining between 32.8% (Cox and Snell R Square) and 48.8% (Nagelkerke R Square) of the variance in reading status, and correctly classified 87.7% of the cases. The odds ratio indicates that for every one unit increase in ORF Expository scores, participants were 1.06 times less likely to be categorized as non-proficient readers on the 4Sight.

Research question 4C. The sixth analysis assessed the predictive values of ORF Expository and RRR Expository with PSSA (See Tables 17 and 18). The full model was not significant, χ^2 (1, *n* = 105) = 1.133, *p* = .287, indicating that RRR Expository did not significantly add to ORF Expository's ability to distinguish between students who scored in the proficient range and students who scored in the non-proficient range on the PSSA (no change in predictive accuracy). The model containing only ORF Expository was statistically significant, χ^2 (1, *n* = 105) = 50.017, *p* < .001, explaining between 37.9% (Cox and Snell R Square) and 55.7% (Nagelkerke R Square) of the variance in reading status, and correctly classified 88.6% of the cases. The odds ratio indicates that for every one unit increase in ORF Expository scores, participants were 1.08 times less likely to be categorized as non-proficient readers on the PSSA.

Research question 5A. The seventh analysis assessed the predictive values of ORF Narrative and Adapted RTF Narrative with GRADE (See Tables 19 and 20). After controlling for ORF Narrative scores, the full model containing Adapted RTF Narrative was statistically significant, χ^2 (1, *n* = 105) = 14.853, *p* < .001, indicating that the model was able to distinguish between students who scored in the proficient range and students who scored in the non-proficient range on the GRADE. This model as a whole explained between 52.2% (Cox and Snell R Square) and 74.8% (Nagelkerke R Square) of the variance in reading status, and correctly classified 89.5% of the cases. Compared to model one which explained between 44.9% and 64.4% of the variance in reading status, correctly classifying 84.8% of the cases, Adapted RTF Narrative increased classification accuracy by 4.7%. Specifically, the addition of Adapted RTF improved (a) sensitivity (Model 1 = 77%; Model 2 = 85%), (b) specificity (Model 1 = 87%; Model 2 = 91%), (c) positive predictive power (Model 1 = 67%; Model 2 = 77%), and (d) negative predictive power (Model 1 = 92%; Model 2 = 95%). The odds ratio indicates that for every one unit increase in Adapted RTF Narrative scores, participants were 1.08 times less likely to be categorized as non-proficient readers on the GRADE. Similarly, for every one unit increase in ORF Narrative scores, participants were 1.11 times less likely be categorized as non-proficient readers on the GRADE.

Research question 5B. The eighth analysis assessed the predictive values of ORF Narrative and Adapted RTF Narrative with 4Sight (See Tables 21 and 22). The full model was not significant, χ^2 (1, *n* = 107) = 0.673, *p* = .412, indicating that Adapted RTF Narrative did not significantly add to ORF Narrative's ability to distinguish between students who scored in the proficient range and students who scored in the non-proficient range on the 4Sight (no change in predictive accuracy). The model containing only ORF Narrative was statistically significant, χ^2 (1, *n* = 107) = 41.808, *p* < .001, explaining between 32.3% (Cox and Snell R Square) and 48.3% (Nagelkerke R Square) of the variance in reading status, and correctly classifying 89.7% of the cases. The odds ratio indicates that for every one unit increase in ORF Narrative scores, participants were 1.06 times less likely to be categorized as non-proficient readers on the 4Sight.

Research question 5C. The ninth analysis assessed the predictive values of ORF Narrative and Adapted RTF Narrative with PSSA (See Tables 23 and 24). After controlling for ORF Narrative scores, the full model containing Adapted RTF Narrative was statistically significant, χ^2 (1, *n* = 106) = 4.989, *p* < .05, indicating that the model was able to distinguish between students who scored in the proficient range and students who scored in the non-proficient range on the PSSA. This model as a whole explained between 43.0% (Cox and Snell R Square) and 63.3% (Nagelkerke R Square) of the variance in reading status, and correctly classified 88.7% of the cases. However, model one explained between 40.2% and 59.3% of the variance in reading status, correctly classifying 89.6% of the cases; consequently, Adapted RTF Narrative *decreased* classification accuracy by 0.9%. In addition, using the Bonferroni corrected alpha level (*p* < .017), Adapted RTF was not identified as a significant predictor (*p* = .038). The odds ratio indicates that for every one unit increase in Adapted RTF Narrative scores, participants were 1.03 times less likely to be categorized as non-proficient readers on the PSSA. Similarly, for every one unit increase in ORF Narrative scores, participants were 1.09 times less likely be categorized as non-proficient readers on the PSSA.

Research question 6A. The tenth analysis assessed the predictive values of ORF Expository and Adapted RTF Expository with GRADE (See Tables 25 and 26). The full model was not significant, χ^2 (1, *n* = 104) = 1.006, *p* = .316, indicating that Adapted RTF Expository did not significantly add to ORF Expository's ability to distinguish between students who scored in the proficient range and students who scored in the non-proficient range on the GRADE (no change in predictive accuracy). The model containing only ORF Expository was statistically significant, χ^2 (1, *n* = 104) = 57.821, *p* < .001, explaining between 42.6% (Cox and Snell R Square) and 61.0% (Nagelkerke R Square) of the variance in reading status, and correctly classified 87.5% of the cases. The odds ratio indicates that for every one unit increase in ORF Expository scores, participants were 1.08 times less likely to be categorized as non-proficient readers on the GRADE.

Research question 6B. The eleventh analysis assessed the predictive values of ORF Expository and Adapted RTF Expository with 4Sight (See Tables 27 and 28). The full model was not significant, χ^2 (1, n = 106) = 0.038, p = .846, indicating that Adapted RTF Expository did not significantly add to ORF Expository's ability to distinguish between students who scored in the proficient range and students who scored in the non-proficient range on the 4Sight (no change in predictive accuracy). The model containing only ORF Expository was statistically significant, χ^2 (1, n = 106) = 42.088, p < .001, explaining between 32.8% (Cox and Snell R Square) and 48.8% (Nagelkerke R Square) of the variance in reading status, and correctly classified 87.7% of the cases. The odds ratio indicates that for every one unit increase in ORF Expository scores, participants were 1.06 times less likely to be categorized as non-proficient readers on the 4Sight.

Research question 6C. The twelfth analysis assessed the predictive values of ORF Expository and Adapted RTF Expository with PSSA (See Tables 29 and 30). The full model was not significant, χ^2 (1, n = 105) = 0.656, p = .418, indicating that Adapted RTF Expository did not significantly add to ORF Expository's ability to distinguish between students who scored in the proficient range and students who scored in the non-proficient range on the PSSA (no change in predictive accuracy). The model containing only ORF Expository was statistically significant, χ^2 (1, n = 105) = 50.017, p < .001, explaining between 37.9% (Cox and Snell R Square) and 55.7% (Nagelkerke R Square) of the variance in reading status, and correctly classified 88.6% of the cases. The odds ratio indicates that for every one unit increase in ORF Expository scores, participants were 1.08 times less likely to be categorized as non-proficient readers on the PSSA.

Chapter Five: Discussion

The purpose of this investigation was to examine the technical adequacy and usability of an oral retelling procedure that employed a rubric scoring method to assess reading comprehension skills of students in third grade. Specifically, this study investigated the convergent and predictive validity of the RRR for identifying children at risk for reading comprehension difficulties on summative reading assessments. The current study aimed to expand on previous investigations of the RRR though direct comparisons of text type within the same study, removal of the passage from view during the retell, addition of item related to sequencing on the expository RRR, and use of logistic regression. Six research questions framed this study. These questions were answered using reading data gathered from 107 elementary school children attending third grade in a public elementary school in Eastern Pennsylvania. Reading data were collected from each participant using CBM (i.e., three narrative and three expository passages with ORF, Adapted RTF, and RRR) and standardized, norm-referenced reading comprehension tests (i.e., GRADE, 4Sight, and PSSA). Pearson product-moment correlations were used to examine the first two research questions, which examined the convergent validity of the RRR, and hierarchical binary logistic regression analyses were used to examine the remaining research questions, which examined the predictive validity of the RRR.

Reliability Analyses

The reliability of the three CBM was assessed using three different techniques. First, alternate form reliability was assessed by examining correlations between the three narrative forms and between the three expository forms of each CBM. Consistent with previous research (e.g., DMG, 2011b), ORF had very strong alternate form reliability for both narrative and expository text. The Adapted RTF measure had strong alternate form reliability; whereas the RRR had moderate alternate form reliability. In contrast to ORF and Adapted RTF in which the only change in assessment across forms is the passage, it is speculated that the RRR had lower alternate form reliability because each passage had a unique scoring template, which provided qualitative answers for each item on the rubric (e.g., setting: stream in an oak forest).

Next, interscorer agreement was determined for the three CBM on a point-by-point basis. Overall, the three CBM demonstrated strong interscorer agreement for both narrative and expository text ($M \geq$ 90% agreement). Further examination of interscorer agreement using the within 2-point criterion for judging interscorer agreement set by Good and Kaminski (2002) yielded acceptable levels of agreement (> 80% of the cases within 2-points of each other) for ORF and RRR and poor agreement for Adapted RTF (43% of Narrative and 56% of Expository cases were within 2-points of each other). Throughout the literature the RTF measure has been criticized for having poor interscorer agreement. In particular, both Bellinger and DiPerna (2011) and Pressley, Hilden, & Shankland (2005) found a significant differences between real time RTF scores and reordered RTF scores.

Notably the range of interscorer agreement was large for RRR (range, 50 to 100% or 0 to 5 points). Revisions to the scoring templates for RRR may improve both alternate form reliability and interscorer agreement. The scoring templates were originally developed using feedback from graduate research assistants. Alternatively, participants' responses from the three studies of the RRR could be used to improve the scoring templates, which in turn may improve interscorer reliability.

Direct Comparison of Text Type

The initial investigation of the narrative version of the RRR was conducted with a different sample than the investigation of the expository version of the RRR; thus, limiting direct comparison of text type. For this study, the same participants were administered the narrative and expository assessments, thereby allowing for direct comparison of text type. Consistent with previous research on the assessment of narrative and expository text (e.g., Best et al., 2008; Diakidoy et al., 2005), this study yielded higher mean scores for the narrative passages across ORF, Adapted RTF, and RRR. Paired-sample t-tests yielded a significant difference between the narrative and expository versions mean score for each CBM. These finding suggest that reading ability and comprehension were influenced by text type.

Early elementary students tend to have greater exposure to narrative text (Graesser, McNamara, & Kulikowich, 2011), resulting in an increased practice with and understanding of narrative store structure. Previous research suggests that comprehension of expository text is often more challenging than narrative text in part due to the influence of a reader’s prior knowledge which is crucial in comprehending expository text to assist the reader in understanding technical vocabulary, generating inferences, and organizing the information to develop a coherent representation of the content (Best et al., 2008; Wolfe & Woodwyk, 2010). Consequently, readers must employ a set of skills when comprehending expository text that are not vital for comprehension of narrative text (Eason, Goldberg, Young, Geist, & Cutting, 2012).

Convergent Validity Analyses

Correlational analysis was used to investigate the convergent validity of the RRR, as outlined in research questions one and two, by comparing performance on the RRR with performance on other established measures of reading comprehension administered at the same point in time. Analysis indicated that all correlations were significant at the $p < .05$ or $p < .01$ levels. Correlation coefficients were interpreted using Evans' (1996) framework (i.e., very weak = 0 to .19, weak = .20 to .39, moderate = .40 to .59, strong = .60 to .79, and very strong = .80 to 1.00). The first two hypotheses regarding the convergent validity of RRR and other measures of reading ability were confirmed and the third hypothesis regarding the influence of text type was not confirmed.

As hypothesized, RRR for narrative and expository text respectively exhibited the strongest relationship with the Adapted RTF. This finding substantiated those from previous investigations of RRR (i.e., Fritschmann et al., 2010; Shapiro et al., 2010), in which the RRR was observed to be moderately correlated with the Adapted RTF, as well as findings from Fuchs et al. (1988) which observed similar methods of scoring oral retell to be highly correlated. The moderate to strong significant correlations between RRR and Adapted RTF found across the three investigations of the RRR indicates that both measures appear to be assessing similar constructs.

The DIBELS 6th edition RTF measure has been criticized for (a) lacking instructional utility and (b) demonstrating weak reliability across scorers (e.g., Bellinger & DiPerna, 2011; Burke & Hagan-Burke, 2007; Marcotte & Hintze, 2009; Pressley et al., 2005; Riedel, 2007). Specifically, the DIBELS 6th edition RTF measure does not provide information about the quality and organization of information the reader amassed from the text, nor does it yield a large sample of comprehension behaviors that can inform instruction and intervention. In contrast, the RRR measure allows for a greater conceptual match to what we know are the key elements of reading comprehension, in particular it allows for evaluation of the accuracy of the components (e.g., narrative text – theme, problem, goal, setting, characters, initiating events, climax, resolution, and ending; expository text – topic, main idea, primary supporting details, and secondary supporting details), sequence, and coherence of the retell (Caldwell & Leslie, 2005). Compared to the DIBELS 6th edition RTF measure, the RRR yields a larger sample of comprehension behaviors that can be useful in making instructional decisions.

Furthermore, consistent with previous investigations of the DIBELS 6th edition RTF measure (e.g., Marcotte & Hintze, 2009), this study found the Adapted RTF measure to demonstrate low levels of interscorer agreement when using the 2-point criterion for judging interscorer agreement set by Good and Kaminski (2002), with only 50% of the interscorer checks within 2-points of one another. In contrast, approximately 90% of the RRR interscorer checks were within 2-points of each other on the final score. Given the weak instructional utility and interscorer reliability of the DIBELS 6th edition RTF measure, the strong significant correlation between RRR and Adapted RTF is important because it suggests that RRR may be a viable alternative to the DIBELS 6th edition RTF measure.

As hypothesized, the RRR for narrative and expository texts respectively exhibited low to moderate correlations (*r* = .23 to .47) with ORF, GRADE, 4Sight, and PSSA. In contrast, ORF was strongly correlated with GRADE, 4Sight, and PSSA. On the surface, one might suppose that the RRR would yield stronger correlations with the

GRADE, 4Sight, and PSSA because they all appear to assess reading comprehension. However, perhaps the low to moderate correlations between RRR and these measures may be attributed to the notion that GRADE, 4Sight, and PSSA measure an individual's *recognition* of words or information, whereas the RRR taps into a reader's *recall* of information (Kucer, 2011). By examining a reader's *recognition* of the correct answer from a list of possible choices, the GRADE, 4Sight, and PSSA are tapping into lower-level comprehension skills (Paris & Paris, 2003). In contrast, by examining a reader's *recall* of information, the RRR is tapping into a deeper-level of understanding (Kintsch, 1998).

The RRR provides information about a reader's understanding of the passage, memory of events, and ability to sequence events and major concepts (Hansen, 1978; Ringler & Weber, 1984). Retelling assessments also allow for observations of metacognitive skills, including a reader's ability to utilize context clues, draw on their prior knowledge, make inferences, monitor their understanding of the text, and employ fix-up strategies to resolve problems with comprehension (e.g., adjust reading speed, look back or forward in text) (Block, 2005; Randi, Grigorenko, & Sternberg, 2005). Retelling (i.e., RRR) is more aligned with authentic reading than multiple-choice tests (i.e., GRADE, 4Sight, and PSSA) (Kintsch, 1998). Retellings have demonstrated consequential validity by having a positive consequence for the student as a result of the experience of *recalling* the text (McKenna & Stahl, 2009). Research has found that practice in retelling improves student's understanding and recognition of narrative and expository story structure elements, ability to recall information, and ability to answer cued recall questions (e.g., Gambrell, Koskinen, & Kapinus, 1991). The RRR may have potential to serve as a formative assessment of reading comprehension. In particular, the RRR could possibly provide real-time information to teachers and students about student understanding of the text. This would allow the teacher and student to respectively adjust teaching and learning while they are still happening.

Noteworthy, correlations between RRR and PSSA were higher (narrative *r* = .46 and expository *r* = .28) in this study than those observed (narrative *r* = .29 and expository *r* = .02) in the previous investigations of the RRR (i.e., Fritschmann et al., 2010; Shapiro et al., 2010). It is speculated that (a) changes to the administration procedures of this study, in particular within students assessment of narrative and expository text, addition of item related to sequencing on the expository RRR, and removal of the passage during retell, as well as (b) the closer proximity of data collection in this study to PSSA testing (i.e., the 2011 PSSA was administered in March, whereas the 2008 PSSA used in the narrative study and the 2009 PSSA used in the expository study were both administered in April) contributed to the higher correlations found between the RRR and PSSA in this study compared to the previous investigations of the RRR.

It was hypothesized that correlations between measures of narrative text would be higher than those between measures of expository text. This hypothesis was based on the notions that (a) narrative texts tend to follow a predictable structure or sequence of events, whereas expository texts tend to have greater structural complexity (Best et al., 2008) and (b) the GRADE, 4Sight, and PSSA include more narrative passages than expository passages. This hypothesis was mostly true for RRR; the correlations between RRR narrative with ORF narrative, GRADE, 4Sight, and PSSA were slightly higher (range, *r* = .36 to .47) than those between RRR expository with ORF expository, GRADE, 4Sight, and PSSA (range, *r* = .23 to .32). However, the reverse was true for Adapted RTF, in which the correlations between Adapted RTF expository and ORF expository, GRADE, 4Sight, and PSSA (range, *r* = .32 to .63) were slightly higher than those between Adapted RTF narrative and ORF narrative, GRADE, 4Sight, and PSSA (range, *r* = .21 to .62). It is speculated that that different findings for the RRR and Adapted RTF are related to differences in the way they assess reading comprehension across narrative and expository text. The content of the RRR for narrative text is different from the content of the RRR for expository text, whereas the Adapted RTF uses the same methodology regardless of text type.

The hypothesis that correlations between measures of narrative text would be higher than those between measures of expository text was not confirmed for ORF; correlations between ORF narrative and expository with GRADE, 4Sight, and PSSA were identical or nearly identical (range, *r* = .72 to .80). The ORF narrative and expository were highly correlated (*r* = .94) suggesting a strong association between scores on ORF narrative and expository. Whereas text type has been shown to influence a student's reading comprehension, the influence of text type is likely to be less prominent for oral reading because if a student is a proficient reader then he or she should be able to apply their reading skills to different texts, thus resulting in proficient oral reading regardless of text type. The strong correlations (range, *r* = .72 to .80) between ORF for narrative and expository text with a measure of reading comprehension (i.e., GRADE) and statewide reading assessments (i.e., 4Sight, and PSSA) are consistent with previous studies about the relationship between ORF and measures of reading comprehension and overall reading ability (e.g., DMG, 2011b; Fuchs et al., 1988; Keller-

Margulis et al., 2008; Reschly et al., 2009; Shapiro et al., 2006; Shapiro et al., 2008; Shinn et al., 1992). In spite of weak face validity, this finding provides further support for ORF as an indicator of overall reading proficiency and reading comprehension. In addition, despite concerns regarding ORF's ability to detect "word callers;" the "word caller" phenomenon may not be as prevalent as teachers may think. Similar to Meisinger et al. (2009) who found low rates of word callers for a sample of third grade students (i.e., approximately 1% of the total sample could be identified as word callers), only two "word callers" were identified in this study (approximately 2% of the total sample). These participants read between 117 and 122 WCPM and performed in the basic range on the PSSA.

Predictive Validity Analyses

To explore the predictive validity of the RRR, as outlined in research questions three through six, twelve hierarchical binary logistic regression analyses were conducted. In each analysis, ORF alone was a statistically significant predictor ($p < .001$) of performance on the criterion measure. This finding confirms that, at third grade, ORF is a powerful metric of overall reading ability. Commensurate with previous research, ORF exhibited strong correlations with measures of overall reading ability ($r = .72$ to $.80$) and emerged as a significant predictor ($p < .001$) of students' performance on standardized reading assessments (i.e., GRADE, 4Sight, and PSSA).

There are four key features of a screening measure: (a) sensitivity – the number of true positives (i.e., the percentage of participants accurately identified by the model as non-proficient readers on the GRADE, 4Sight, or PSSA), (b) specificity – the number of true negatives (i.e., the percentage of participants accurately identified by the model as proficient readers on the GRADE, 4Sight, or PSSA), (c) positive predictive power – the percentage of cases with "positive" test results who are correctly diagnosed with the ailment (i.e., the percentage of cases that the model predicted to be non-proficient readers on the GRADE, 4Sight, or PSSA and were actually non-proficient readers on the GRADE, 4Sight, or PSSA), and (d) negative predictive power – the percentage of cases with "negative" test results who are correctly diagnosed as *not* possessing the ailment (i.e., the percentage of cases the model predicted to be proficient readers on the GRADE, 4Sight, or PSSA and were actually proficient readers on the GRADE, 4Sight, or PSSA). A perfect screening measure would have 100% positive predictive power and negative predictive power; however, error is inherent in every measure. Therefore, acceptable accuracy of a screening measure comes down to a trade-off between sensitivity and specificity (Jenkins et al., 2007). More weight is given to sensitivity because false negatives are a far more egregious type of error than false positives, because false negatives would deny access to intervention for students who most needed it, whereas false positives would expend resources on students who do not require intervention (Jenkins et al., 2007).

Jenkins and colleagues (2007) recommended a minimum acceptable level of sensitivity of 90%, which corresponds to a false negative rate of 10%. In this investigation, the sensitivity of ORF alone ranged from 77% to 87%, which translates to a false negative rate of ≤ 23%, meaning that approximately six students were misidentified as proficient readers. Compton and colleagues (2010) recommended a minimum acceptable level of specificity of 80%. The specificity of ORF ranged from 87% to 91%, which translates to a false positive rate of ≤ 13%, meaning that approximately ten students were misidentified as non-proficient readers. Three studies have investigated the sensitivity and specificity of ORF in predicting performance on the PSSA. Shapiro et al. (2008) found greater sensitivity (range, .79 to .97) and weaker specificity (range, .49 to .61) for ORF scores predicting performance on the PSSA, whereas Keller-Margulis et al. (2008) found weaker sensitivity (range, .71 to .77) and greater specificity (range, .78 to .90) and Shapiro et al. (2006) found greater sensitivity for one district (sensitivity range, .70 to .76; specificity range, .70 to .75) and greater specificity for another district (sensitivity range, .69 to .86; specificity range .67 to .83). Taken together, these findings along with those from the current study indicate that while ORF is not a perfect metric, it has technical adequacy and utility as a screening tool at third grade to identify those students at risk for poor performance on state reading assessments.

It was hypothesized that RRR for narrative and expository text respectively would add significantly to ORF for narrative and expository text's accurate prediction of non-proficient and proficient readers on the GRADE, 4Sight, and PSSA. Of the six analyses that included the RRR, the RRR narrative was identified as a statistically significant predictor in two of the analyses (GRADE *p* = .001; PSSA *p* = .025). However, using the Bonferroni corrected alpha level (*p* < .017), the RRR narrative was only a statistically significant predictor for the GRADE. The RRR expository was not identified as a statistically significant predictor for any of the analyses. The GRADE, 4Sight, and PSSA have a greater concentration of narrative text, which may explain why the RRR expository was not identified as a statistically significant predictor of the dependent variables.

It was also hypothesized that the combination of RRR and ORF would be a stronger prediction model than the combination of Adapted RTF and ORF. Of the six analyses that included the Adapted RTF, the Adapted RTF narrative was identified as a significant predictor in the same two analyses as RRR narrative (GRADE p = .002; PSSA p = .038). However, using the Bonferroni corrected alpha level, the Adapted RTF narrative was only a significant predictor for the GRADE. In addition, Adapted RTF narrative was found to decrease ORF's prediction of the PSSA.

Overall, the predictive validity hypotheses were not confirmed. This may be attributed to the strength of ORF as a predictor of reading proficiency at third grade. This finding may also be attributed to weak power. An a priori power analysis was conducted using Power Analysis and Sample Size software (PASS; Hintze, 2008). PASS determined the number of study participants using Hsieh, Bloch, and Larsen's (1998) method for calculating sample size for logistic regression. Results indicated that Logistic regression of a binary dependent variable and two continuous independent variables required a sample size of 103 to achieve 80% power at a 0.05 significance level to detect a change in Probability (Y = 1) from the value of 0.250 (P0) at the mean of X to 0.150 (P1) when X is increased to one and half standard deviations above the mean, resulting in an odds ratio of 0.333 (Hintze; Hsieh et al.). All analyses included a sample ≥ 104. However, post-hoc power analysis (See Table 31) revealed weak power. With the exception of the first logistic regression analysis (RQ3A; Power = 89%), the RRR's average power was 26% (range, 13% to 42%) and Adapted RTF's average power was 3.5% (range, 3% to 5%). The different findings between the a priori and post-hoc power analyses are due to differences in the odds ratio. The a priori power analysis used the standard/default setting for the odds ratio (0.529). However, the odds ratio in the actual data was much higher for both RRR (range, 0.657 to 0.773) and Adapted RTF (range, .934 to 1.011), thus requiring a much larger sample size to detect significance. Consequently, weak power may have partially contributed to the non-significant contribution of RRR and Adapted RTF. Difference in scaling for RRR may also have contributed to the size of the odds ratio and subsequent power to detect significance. The RRR is based on a 10-point scale whereas ORF and Adapted RTF approximately range from 0 to 200 words; thus, if a participant increases their score by 1 point it has a great impact for RRR's odds ratio than if a participant reads/retells an additional word for ORF/Adapted RTF.

Limitations and Directions for Future Research

There are several limitations to this study. Most remarkable, the sample size was not large enough to detect significance in the predictive validity analyses. Future research should seek to sample from a much larger population; which may require aggregating data from students across multiple schools. An additional limitation to this study's sample was that the participants were not from diverse backgrounds, as the majority of participants were Caucasian; therefore, limiting generalizability of the findings. Future research may seek to improve generalizability by sampling from a more diverse population and/or examining the RRR with individuals from linguistically diverse backgrounds, whose first language is not English. Cultural and linguistic differences might impact students' performance on the RRR due to the strong reliance on oral language processing (Snyder, Caccamise, & Wise, 2005). This study included only third grade students. Instruction in school shifts from narrative to expository text as a function of grade, with third through fifth grade marking a critical transition period as students shift from narrative to expository text and shift from "learning to read" to "reading to learn" (Graesser et al., 2011; Stevens, Slavin, & Farnish, 1991). Consequently, future research may seek to investigate the contextual appropriateness, technical adequacy, and usability of the RRR narrative and expository versions with students in lower elementary grades compared to students in higher elementary grades. In particular, this will be helpful in determining whether the RRR narrative and expository versions are relevant for specific grades or if they cut across the grades.

RRR did not strongly or significantly add to ORF's prediction of the criterion-measures of reading proficiency. Consequently, RRR did not demonstrate the technical adequacy required for identification as a universal screening measure. Perhaps RRR would be better suited as diagnostic tool as opposed to a universal screener. Whereas ORF possesses the technical adequacy and usability as a universal screener, it lacks content validity for designing interventions for students who struggle with comprehension. ORF only yields information regarding a reader's speed, accuracy, and prosody of oral production of text, and does not directly provide any information regarding a reader's comprehension of the text. Future research on the RRR may seek to investigate the use of RRR within a multiple-gated screening process (See Figure 1) to explore RRR's technical adequacy and usability in making instructional decisions. Using a multiple-gated screening process may help to improve sensitivity and specificity of ORF, reducing the number of false negatives and false positives to more acceptable ranges. The RRR requires staff training to administer and score, as well as additional administration time, thus, using RRR as a diagnostic tool within a multiple-gated screening process would also conserve resources.

This study required the RRR to be audio recorded and scored at a later time. This method is not feasible for school settings in terms of resources and time. Future research should examine whether RRR can be scored live through examining the interscorer reliability between scores obtained via the live scoring and audio recorded scoring.

Conclusions

This investigation reinforces the strength of ORF at third grade in predicting reading ability on summative assessments. At third grade, ORF is a powerful metric of overall reading ability. Findings showed that this brief (i.e., 1-minute sample of reading) had strong predictive power to identify third grade students who will likely demonstrate proficient or non-proficient reading ability on criterion-measures of reading comprehension and state reading assessments. Results from this study add to the extensive literature base which supports the technical adequacy and utility of the ORF as a powerful screening measure at third grade.

Although a strong measure, ORF is not a perfect metric in that it (a) yielded unacceptable rates of false negatives and false positives, (b) lacks the content validity for designing instruction to address specific deficits in reading comprehension, and (c) lacks face validity as an assessment of reading comprehension. It has been suggested that adding a retell assessment to ORF would improve the predictive, content, and face validity of the assessment. A retell measure provides different information about students' reading ability than ORF. Whereas ORF only yields information regarding a reader's speed, accuracy, and prosody of oral production of text, a retell assessment can provide information about the quantity, quality, and organization of information a reader ascertained from reading the text (Winograd et al. 1989), which can be useful in identifying individuals who struggle with comprehension (i.e., predictive validity) and informing instruction and intervention development (i.e., content validity). In addition, retell has greater face validity than ORF as an assessment of reading comprehension. In particular, Reed and Petscher (in press) found middle school teachers to rate a retell measure more favorably than an ORF measure; teachers doubted the ORF measure as an indicator of students' comprehension skills and indicated that the retell measure provided more valuable information than ORF because it was more akin to classroom instruction. Reed and Petscher also found teachers to view ORF as more acceptable when combined with retell.

The current study investigated whether the RRR could be added to ORF to improve the predictive, content, and face validity of the assessment. Findings showed that ORF alone was a strong predictor of overall reading ability in third grade. More research is needed, particularly with a larger sample, to determine the usability of the RRR. Notably the RRR was able to detect the two "Word Callers" in this study. These students exhibited ORF between 117 to 122 WCPM, performed in the basic range on the PSSA, and earned less than 5 points on the RRR. Perhaps, the RRR would be better suited as a diagnostic tool (instead of a universal screening tool) within a multiple-gated screening process. Findings from the current study suggest that the RRR may be a viable alternative to the DIBELS 6th edition RTF measure. The DIBELS 6th edition RTF measure only provides information about quantity of information a reader amassed from the text, thus yielding a smaller sample of comprehension behaviors than the RRR measure. In particular, the RRR measure provides a greater conceptual match to what we know are the key elements of reading comprehension, including the accuracy of components, sequence, coherence of the retell (Caldwell & Leslie, 2005). Consequently, the RRR measure also possesses greater face validity as a measure of reading comprehension than the RTF measure. The RRR also has greater interscorer reliability than the RTF. In sum, the RRR may have potential to better inform, than RTF, formative and diagnostic assessments of students reading comprehension; however, further investigation is needed to establish RRR's utility as a CBM of reading comprehension.

Table 1

Demographic Characteristics of the Elementary School.

Characteristics	Elementary School
Locale	Suburban
Size	556 students in Grades 3-6
Gender	
Male	292
Female	264
Race/Ethnicity	
White	475
Latino/Hispanic	50
Multiracial	12
Black	11
Students with IEPs	108
Students who are Economically Disadvantaged	246
AYP Status	Made
School's Overall Results in Reading, All Students	69% of students proficient/advanced
School's Overall Results in Mathematics, All Students	77% of students proficient/advanced

Note. Data compiled from Pennsylvania Department of Education (2011). *Academic Achievement Report: 2010-2011.* Available at: http://paayp.emetric.net/.

Table 2

Characteristics of the Narrative and Expository Passages.

Measures	Narrative Passages			Expository Passages		
	Sofa	Fish	Party	Giraffe	Flamingo	Owl
Number of Words	275	283	271	313	294	258
Number of Sentences	28	28	23	35	34	30
Spache Readability	3.4	3.3	3.6	2.9	3.2	3.6
Lexile	660	610	660	570	610	600

Table 3

Descriptive Statistics for All Variables

Variable	*n*	M	SD	Skewness	Kurtosis	Minimum	Maximum
ORF Narrative	107	99.42	32.09	-.25	-.46	24	166
Adapted RTF Narrative	107	69.98	25.35	.09	-.35	13	128
RRR Narrative	107	6.16	1.77	-.29	-.69	2	9
ORF Expository	107	96.21	30.99	-.68	-.11	15	145
Adapted RTF Expository	107	49.55	22.05	.75	1.43	11	139
RRR Expository	107	5.66	1.65	-.01	.03	2	10
GRADE Comprehension	105	34.50	8.80	-1.01	.17	11	46
4Sight Reading	107	20.40	6.27	-.75	-.13	3	30
PSSA Reading	106	31.34	9.23	-.76	-.29	9	44

Note. ORF = Oral Reading Fluency; RTF = Retell Fluency; RRR = Reading Retell Rubric; GRADE Comprehension = Group Reading Assessment and Diagnostic Evaluation Comprehension Composite; 4Sight Reading = Pennsylvania 4Sight Reading Benchmark Assessment; PSSA Reading = Pennsylvania System of Student Assessment Reading Assessment.

Table 4

Pearson Correlations between All Variables

Variable	*n*	1	2	3	4	5	6	7	8	9
1. ORF Narrative	106	—								
2. Adapted RTF Narrative	106	.26**	—							
3. RRR Narrative	106	.36**	.62**	—						
4. ORF Expository	106	.94**	.20*	.33**	—					
5. Adapted RTF Expository	106	.34**	.62**	.50**	.34**	—				
6. RRR Expository	106	.23*	.41**	.40**	.23*	.63**	—			
7. GRADE Comprehension	104	.80**	.32**	.47**	.80**	.38**	.32**	—		
8. 4Sight Reading	105	.72**	.21*	.37**	.72**	.32**	.31**	.81**	—	
9. PSSA Reading	105	.76**	.34**	.46**	.74**	.36**	.28**	.88**	.80**	—

Note. ORF = Oral Reading Fluency; RTF = Retell Fluency; RRR = Reading Retell Rubric; GRADE Comprehension = Group Reading Assessment and Diagnostic Evaluation Comprehension Composite; 4Sight Reading = Pennsylvania 4Sight Reading Benchmark Assessment; PSSA Reading = Pennsylvania System of Student Assessment Reading Assessment.
* *p* < .05. ** *p* < .01.

Table 5

Frequency of Reading Classification across the Dependent Variables

Variable	Proficient Readers	Non-Proficient Readers	Total
Narrative Text			
GRADE Comprehension	75	30	105
4Sight Reading	81	26	107
PSSA Reading	79	27	106
Expository Text			
GRADE Comprehension	74	30	104
4Sight Reading	80	26	106
PSSA Reading	78	27	105

Note. GRADE Comprehension = Group Reading Assessment and Diagnostic Evaluation Comprehension Composite; 4Sight Reading = Pennsylvania 4Sight Reading Benchmark Assessment; PSSA Reading = Pennsylvania System of Student Assessment Reading Assessment.

Table 6

Summary of Findings across the Twelve Hierarchical Binary Logistic Regression Analyses for ORF with RRR or RTF Predicting GRADE, 4Sight, or PSSA

Dependent Variable	Block	Predictor	*B*	*SE*	Wald statistic	*p*	Classification Accuracy			
							Sens.	Spec.	PPP	NPP
GRADE										
Narrative	1	ORF N	-.09	.02	23.12	< .001	.77	.87	.67	.92
RQ3A	2	ORF N	-.10	.02	17.65	< .001	.88	.90	.73	.96
		RRR N	-.77	.23	10.92	.001				
RQ5A	2	ORF N	-.11	.02	19.80	< .001	.85	.91	.77	.95
		RTF N	-.07	.02	9.79	.002				
Expository	1	ORF E	-.08	.02	24.88	< .001	.87	.88	.67	.96
RQ4A	2	ORF E	-.08	.02	23.30	< .001	.84	.89	.70	.95
		RRR E	-.31	.19	2.63	.105				
RQ6A	2	ORF E	-.08	.02	22.92	< .001	.83	.88	.67	.95
		RTF E	-.02	.02	0.98	.322				
4Sight										
Narrative	1	ORF N	-.06	.01	22.52	< .001	.86	.91	.69	.96
RQ3B	2	ORF N	-.06	.01	19.62	< .001	.90	.91	.69	.98
		RRR N	-.26	.17	2.38	.123				
RQ5B	2	ORF N	-.06	.01	21.35	< .001	.86	.91	.69	.96
		RTF N	.01	.01	0.66	.415				
Expository	1	ORF E	-.06	.01	24.07	< .001	.81	.89	.65	.95
RQ4B	2	ORF E	-.06	.01	22.38	< .001	.82	.90	.69	.95
		RRR E	-.33	.19	3.04	.081				
RQ6B	2	ORF E	-.06	.01	21.98	< .001	.78	.90	.69	.94
		RTF E	-.00	.02	0.04	.847				
PSSA										
Narrative	1	ORF N	-.08	.02	23.28	< .001	.86	.90	.70	.96
RQ3C	2	ORF N	-.08	.02	20.48	< .001	.91	.92	.74	.97
		RRR N	-.42	.19	5.03	.025				
RQ5C	2	ORF N	-.08	.02	22.58	< .001	.83	.90	.70	.95
		RTF N	-.04	.02	4.30	.038				
Expository	1	ORF E	-.07	.01	24.94	< .001	.83	.90	.70	.95
RQ4C	2	ORF E	-.07	.01	23.80	< .001	.83	.90	.70	.95
		RRR E	-.19	.18	1.09	.296				
RQ6C	2	ORF E	-.07	.01	22.62	< .001	.83	.90	.70	.95
		RTF E	-.01	.02	0.65	.422				

Note. GRADE = Group Reading Assessment and Diagnostic Evaluation Comprehension Composite; 4Sight = Pennsylvania 4Sight Reading Benchmark Assessment; PSSA = Pennsylvania System of Student Assessment Reading Assessment; RQ = Research Question; ORF = Oral Reading Fluency; RTF = Adapted Retell Fluency; RRR = Reading Retell Rubric; N = Narrative; E = Expository. Bonferoni adjusted $p < .017$.

Table 7

Summary of Hierarchical Binary Logistic Regression Analysis for ORF and RRR Narrative Text Predicting GRADE

Predictor	*B*	*SE*	*OR*	95% CI	Wald statistic	*p*
Block 1						
ORF Narrative	-.09	.02	.91	[0.88, 0.95]	23.12	<.001
Block 2						
ORF Narrative	-.10	.02	.91	[0.86, 0.95]	17.65	<.001
RRR Narrative	-.77	.23	.46	[0.29, 0.73]	10.92	.001

Note. ORF = Oral Reading Fluency; RRR = Reading Retell Rubric; GRADE = Group Reading Assessment and Diagnostic Evaluation Comprehension Composite; CI = confidence interval for odds ratio (OR). Bonferroni adjusted *p* < .017.

Table 8

Classification Table for ORF and RRR Narrative Text Predicting GRADE

	Predicted group		
Actual group	Proficient	Non-Proficient	Percentage Correct
Block 1			
Proficient	69	6	92.0
Non-Proficient	10	20	66.7
Overall			84.8
Block 2			
Proficient	72	3	96.0
Non-Proficient	8	22	73.3
Overall			89.5

Note. ORF = Oral Reading Fluency; RRR = Reading Retell Rubric; GRADE = Group Reading Assessment and Diagnostic Evaluation Comprehension Composite.

Table 9

Summary of Hierarchical Binary Logistic Regression Analysis for ORF and RRR Narrative Text Predicting 4Sight

Predictor	*B*	*SE*	*OR*	95% CI	Wald statistic	*p*
Block 1						
ORF Narrative	-.06	.01	.94	[0.92, 0.97]	22.52	<.001
Block 2						
ORF Narrative	-.06	.01	.94	[0.92, 0.97]	19.62	<.001
RRR Narrative	-.26	.17	.77	[0.56, 1.07]	2.38	.123

Note. ORF = Oral Reading Fluency; RRR = Reading Retell Rubric; 4Sight = Pennsylvania 4Sight Reading Benchmark Assessment; CI = confidence interval for odds ratio (OR). Bonferroni adjusted $p < .017$.

Table 10

Classification Table for ORF and RRR Narrative Text Predicting 4Sight

	Predicted group		
Actual group	Proficient	Non-Proficient	Percentage Correct
Block 1			
Proficient	78	3	96.3
Non-Proficient	8	18	69.2
Overall			89.7
Block 2			
Proficient	79	2	97.5
Non-Proficient	8	18	69.2
Overall			90.7

Note. ORF = Oral Reading Fluency; RRR = Reading Retell Rubric; 4Sight = Pennsylvania 4Sight Reading Benchmark Assessment.

Table 11

Summary of Hierarchical Binary Logistic Regression Analysis for ORF and RRR Narrative Text Predicting PSSA

Predictor	*B*	*SE*	*OR*	95% CI	Wald statistic	*p*
Block 1						
ORF Narrative	-.08	.02	.92	[0.90, 0.95]	23.28	<.001
Block 2						
ORF Narrative	-.08	.02	.93	[0.90, 0.96]	20.48	<.001
RRR Narrative	-.42	.19	.66	[0.46, 0.95]	5.03	.025

Note. ORF = Oral Reading Fluency; RRR = Reading Retell Rubric; PSSA = Pennsylvania System of Student Assessment Reading Assessment; CI = confidence interval for odds ratio (OR). Bonferroni adjusted *p* < .017.

Table 12

Classification Table for ORF and RRR Narrative Text Predicting PSSA

	Predicted group		
Actual group	Proficient	Non-Proficient	Percentage Correct
Block 1			
Proficient	76	3	96.2
Non-Proficient	8	19	70.4
Overall			89.6
Block 2			
Proficient	77	2	97.5
Non-Proficient	7	20	74.1
Overall			91.5

Note. ORF = Oral Reading Fluency; RRR = Reading Retell Rubric; PSSA = Pennsylvania System of Student Assessment Reading Assessment.

Table 13

Summary of Hierarchical Binary Logistic Regression Analysis for ORF and RRR Expository Text Predicting GRADE

Predictor	*B*	*SE*	*OR*	95% CI	Wald statistic	*p*
Block 1						
ORF Expository	-.08	.02	.93	[0.90, 0.95]	24.88	<.001
Block 2						
ORF Expository	-.08	.02	.93	[0.90, 0.96]	23.30	<.001
RRR Expository	-.31	.19	.73	[0.50, 1.07]	2.63	.105

Note. ORF = Oral Reading Fluency; RRR = Reading Retell Rubric; GRADE = Group Reading Assessment and Diagnostic Evaluation Comprehension Composite; CI = confidence interval for odds ratio (OR). Bonferroni adjusted *p* < .017.

Table 14

Classification Table for ORF and RRR Expository Text Predicting GRADE

	Predicted group		
Actual group	Proficient	Non-Proficient	Percentage Correct
Block 1			
Proficient	71	3	95.9
Non-Proficient	10	20	66.7
Overall			87.5
Block 2			
Proficient	70	4	94.6
Non-Proficient	9	21	70.0
Overall			87.5

Note. ORF = Oral Reading Fluency; RRR = Reading Retell Rubric; GRADE = Group Reading Assessment and Diagnostic Evaluation Comprehension Composite.

Table 15

Summary of Hierarchical Binary Logistic Regression Analysis for ORF and RRR Expository Text Predicting 4Sight

Predictor	*B*	*SE*	*OR*	95% CI	Wald statistic	*p*
Block 1						
ORF Expository	-.06	.01	.94	[0.92, 0.97]	24.07	<.001
Block 2						
ORF Expository	-.06	.01	.94	[0.92, 0.97]	22.38	<.001
RRR Expository	-.33	.19	.72	[0.50, 1.04]	3.04	.081

Note. ORF = Oral Reading Fluency; RRR = Reading Retell Rubric; 4Sight = Pennsylvania 4Sight Reading Benchmark Assessment; CI = confidence interval for odds ratio (OR). Bonferroni adjusted *p* < .017.

Table 16

Classification Table for ORF and RRR Expository Text Predicting 4Sight

	Predicted group		
Actual group	Proficient	Non-Proficient	Percentage Correct
Block 1			
Proficient	76	4	95.0
Non-Proficient	9	17	65.4
Overall			87.7
Block 2			
Proficient	76	4	95.0
Non-Proficient	8	18	69.2
Overall			88.7

Note. ORF = Oral Reading Fluency; RRR = Reading Retell Rubric; 4Sight = Pennsylvania 4Sight Reading Benchmark Assessment.

Table 17

Summary of Hierarchical Binary Logistic Regression Analysis for ORF and RRR Expository Text Predicting PSSA

Predictor	*B*	*SE*	*OR*	95% CI	Wald statistic	*p*
Block 1						
ORF Expository	-.07	.01	.93	[0.91, 0.96]	24.94	<.001
Block 2						
ORF Expository	-.07	.01	.94	[0.91, 0.96]	23.80	<.001
RRR Expository	-.19	.18	.83	[0.58, 1.18]	1.09	.296

Note. ORF = Oral Reading Fluency; RRR = Reading Retell Rubric; PSSA = Pennsylvania System of Student Assessment Reading Assessment; CI = confidence interval for odds ratio (OR). Bonferroni adjusted *p* < .017.

Table 18

Classification Table for ORF and RRR Expository Text Predicting PSSA

	Predicted group		
Actual group	Proficient	Non-Proficient	Percentage Correct
Block 1			
Proficient	74	4	94.9
Non-Proficient	8	19	70.4
Overall			88.6
Block 2			
Proficient	74	4	94.9
Non-Proficient	8	19	70.4
Overall			88.6

Note. ORF = Oral Reading Fluency; RRR = Reading Retell Rubric; PSSA = Pennsylvania System of Student Assessment Reading Assessment.

Table 19

Summary of Hierarchical Binary Logistic Regression Analysis for ORF and Adapted RTF Narrative Text Predicting GRADE

Predictor	*B*	*SE*	*OR*	95% CI	Wald statistic	*p*
Block 1						
ORF Narrative	-.09	.02	.91	[0.88, 0.95]	23.12	<.001
Block 2						
ORF Narrative	-.11	.02	.90	[0.86, 0.94]	19.80	<.001
Adapted RTF Narrative	-.07	.02	.93	[0.89, 0.98]	9.79	.002

Note. ORF = Oral Reading Fluency; RTF = Retell Fluency; GRADE = Group Reading Assessment and Diagnostic Evaluation Comprehension Composite; CI = confidence interval for odds ratio (OR). Bonferroni adjusted *p* < .017.

Table 20

Classification Table for ORF and Adapted RTF Narrative Text Predicting GRADE

	Predicted group		
Actual group	Proficient	Non-Proficient	Percentage Correct
Block 1			
Proficient	69	6	92.0
Non-Proficient	10	20	66.7
Overall			84.8
Block 2			
Proficient	71	4	94.7
Non-Proficient	7	23	76.7
Overall			89.5

Note. ORF = Oral Reading Fluency; RTF = Retell Fluency; GRADE = Group Reading Assessment and Diagnostic Evaluation Comprehension Composite.

Table 21

Summary of Hierarchical Binary Logistic Regression Analysis for ORF and Adapted RTF Narrative Text Predicting 4Sight

Predictor	*B*	*SE*	*OR*	95% CI	Wald statistic	*p*
Block 1						
ORF Narrative	-.06	.01	.94	[0.92, 0.97]	22.52	<.001
Block 2						
ORF Narrative	-.06	.01	.94	[0.91, 0.96]	21.35	<.001
Adapted RTF Narrative	.01	.01	1.01	[0.99, 1.04]	0.66	.415

Note. ORF = Oral Reading Fluency; RTF = Retell Fluency; 4Sight = Pennsylvania 4Sight Reading Benchmark Assessment; CI = confidence interval for odds ratio (OR). Bonferroni adjusted p < .017.

Table 22

Classification Table for ORF and Adapted RTF Narrative Text Predicting 4Sight

	Predicted group		
Actual group	Proficient	Non-Proficient	Percentage Correct
Block 1			
Proficient	78	3	96.3
Non-Proficient	8	18	69.2
Overall			89.7
Block 2			
Proficient	78	3	96.3
Non-Proficient	8	18	69.2
Overall			89.7

Note. ORF = Oral Reading Fluency; RTF = Retell Fluency; 4Sight = Pennsylvania 4Sight Reading Benchmark Assessment.

Table 23

Summary of Hierarchical Binary Logistic Regression Analysis for ORF and Adapted RTF Narrative Text Predicting PSSA

Predictor	*B*	*SE*	*OR*	95% CI	Wald statistic	*p*
Block 1						
ORF Narrative	-.08	.02	.92	[0.90, 0.95]	23.28	<.001
Block 2						
ORF Narrative	-.08	.02	.92	[0.89, 0.95]	22.58	<.001
Adapted RTF Narrative	-.04	.02	.97	[0.93, 1.00]	4.30	.038

Note. ORF = Oral Reading Fluency; RTF = Retell Fluency; PSSA = Pennsylvania System of Student Assessment Reading Assessment; CI = confidence interval for odds ratio (OR). Bonferroni adjusted *p* < .017.

Table 24

Classification Table for ORF and Adapted RTF Narrative Text Predicting PSSA

	Predicted group		
Actual group	Proficient	Non-Proficient	Percentage Correct
Block 1			
Proficient	76	3	96.2
Non-Proficient	8	19	70.4
Overall			89.6
Block 2			
Proficient	75	4	94.9
Non-Proficient	8	19	70.4
Overall			88.7

Note. ORF = Oral Reading Fluency; RTF = Retell Fluency; PSSA = Pennsylvania System of Student Assessment Reading Assessment.

Table 25

Summary of Hierarchical Binary Logistic Regression Analysis for ORF and Adapted RTF Expository Text Predicting GRADE

Predictor	*B*	*SE*	*OR*	95% CI	Wald statistic	*p*
Block 1						
ORF Expository	-.08	.02	.93	[0.90, 0.95]	24.88	<.001
Block 2						
ORF Expository	-.08	.02	.93	[0.90, 0.96]	22.92	<.001
Adapted RTF Expository	-.02	.02	.98	[0.95, 1.02]	0.98	.322

Note. ORF = Oral Reading Fluency; RTF = Retell Fluency; GRADE = Group Reading Assessment and Diagnostic Evaluation Comprehension Composite; CI = confidence interval for odds ratio (OR). Bonferroni adjusted *p* < .017.

Table 26

Classification Table for ORF and Adapted RTF Expository Text Predicting GRADE

	Predicted group		
Actual group	Proficient	Non-Proficient	Percentage Correct
Block 1			
Proficient	71	3	95.9
Non-Proficient	10	20	66.7
Overall			87.5
Block 2			
Proficient	70	4	94.6
Non-Proficient	10	20	66.7
Overall			86.5

Note. ORF = Oral Reading Fluency; RTF = Retell Fluency; GRADE = Group Reading Assessment and Diagnostic Evaluation Comprehension Composite.

Table 27

Summary of Hierarchical Binary Logistic Regression Analysis for ORF and Adapted RTF Expository Text Predicting 4Sight

Predictor	*B*	*SE*	*OR*	95% CI	Wald statistic	*p*
Block 1						
ORF Expository	-.06	.01	.94	[0.92, 0.97]	24.07	<.001
Block 2						
ORF Expository	-.06	.01	.94	[0.92, 0.97]	21.98	<.001
Adapted RTF Expository	-.00	.02	1.00	[0.97, 1.03]	0.04	.847

Note. ORF = Oral Reading Fluency; RTF = Retell Fluency; 4Sight = Pennsylvania 4Sight Reading Benchmark Assessment; CI = confidence interval for odds ratio (OR). Bonferroni adjusted *p* < .017.

Table 28

Classification Table for ORF and Adapted RTF Expository Text Predicting 4Sight

	Predicted group		
Actual group	Proficient	Non-Proficient	Percentage Correct
Block 1			
Proficient	76	4	95.0
Non-Proficient	9	17	65.4
Overall			87.7
Block 2			
Proficient	75	5	93.8
Non-Proficient	8	18	69.2
Overall			87.7

Note. ORF = Oral Reading Fluency; RTF = Retell Fluency; 4Sight = Pennsylvania 4Sight Reading Benchmark Assessment.

Table 29

Summary of Hierarchical Binary Logistic Regression Analysis for ORF and Adapted RTF Expository Text Predicting PSSA

Predictor	*B*	*SE*	*OR*	95% CI	Wald statistic	*p*
Block 1						
ORF Expository	-.07	.01	.93	[0.91, 0.96]	24.94	<.001
Block 2						
ORF Expository	-.07	.01	.94	[0.91, 0.96]	22.62	<.001
Adapted RTF Expository	-.01	.02	.99	[0.95, 1.02]	0.65	.422

Note. ORF = Oral Reading Fluency; RTF = Retell Fluency; PSSA = Pennsylvania System of Student Assessment Reading Assessment; CI = confidence interval for odds ratio (OR). Bonferroni adjusted *p* < .017.

Table 30

Classification Table for ORF and Adapted RTF Expository Text Predicting PSSA

	Predicted group		
Actual group	Proficient	Non-Proficient	Percentage Correct
Block 1			
Proficient	74	4	94.9
Non-Proficient	8	19	70.4
Overall			88.6
Block 2			
Proficient	74	4	94.9
Non-Proficient	8	19	70.4
Overall			88.6

Note. ORF = Oral Reading Fluency; RTF = Retell Fluency; PSSA = Pennsylvania System of Student Assessment Reading Assessment.

Table 31

Post-Hoc Power Analysis Results for Hierarchical Binary Logistic Regression analyses

Analysis	*n*	Power	*p*	P0	P1	OR	R Squared	β
RRR Narrative								
RQ3A: GRADE	105	89%	.05	.250	.133	.462	.129	.109
RQ3B: 4Sight	107	19%	.05	.250	.205	.773	.129	.812
RQ3C: PSSA	106	42%	.05	.250	.180	.657	.129	.585
RRR Expository								
RQ4A: GRADE	104	27%	.05	.250	.196	.732	.053	.733
RQ4B: 4Sight	106	30%	.05	.250	.193	.719	.053	.703
RQ4C: PSSA	105	13%	.05	.250	.216	.826	.053	.872
RTF Narrative								
RQ5A: GRADE	105	5%	.05	.250	.237	.934	.068	.952
RQ5B: 4Sight	107	3%	.05	.250	.252	1.011	.068	.972
RQ5C: PSSA	106	4%	.05	.250	.244	.966	.068	.965
RTF Expository								
RQ6A: GRADE	104	3%	.05	.250	.247	.983	.113	.971
RQ6B: 4Sight	106	3%	.05	.250	.249	.997	.113	.974
RQ6C: PSSA	105	3%	.05	.250	.247	.986	.113	.971

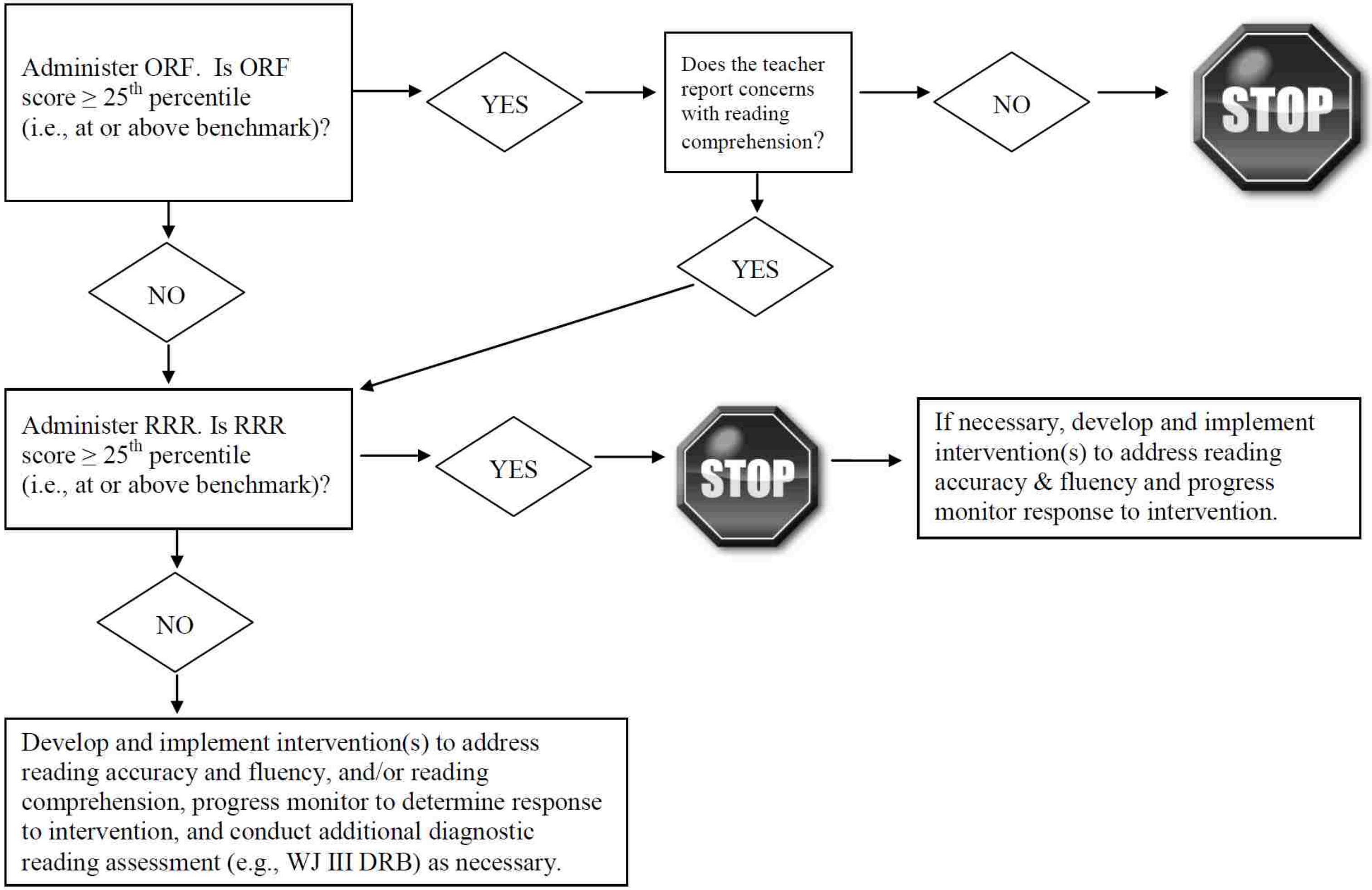

Administer ORF. Is ORF score ≥ 25th percentile (i.e., at or above benchmark)?
YES
Does the teacher report concerns with reading comprehension?
NO
STOP
YES
NO
Administer RRR. Is RRR score ≥ 25th percentile (i.e., at or above benchmark)?
YES
STOP
If necessary, develop and implement intervention(s) to address reading accuracy & fluency and progress monitor response to intervention.
NO
Develop and implement intervention(s) to address reading accuracy and fluency, and/or reading comprehension, progress monitor to determine response to intervention, and conduct additional diagnostic reading assessment (e.g., WJ III DRB) as necessary.

Figure 1. Example of a multiple-gated screening process using the Reading Retell Rubric.

References

Adams, M. J. (1990). *Beginning to read.* Cambridge, MA: MIT Press.

Adams, M. J., Foorman, B. R., Lundberg, I., & Beeler, T. (1998). *Phonemic awareness in young children.* Baltimore: Paul H. Brookes.

American Educator. (1995). *Learning to read: Schooling's first mission, 19*(2), 3–6.

Baker, L., & Brown, A. L. (1984). Metacognitive skills and reading. In P. D. Pearson, R. Barr, M. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (Vol. 1) (pp. 353–394). New York: Longman.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48,* 5–37.

Barger, J. (2003). *Comparing the DIBELS oral reading fluency indicator and the North Carolina end of grade reading assessment* (Technical Report). Asheville, NC: North Carolina Teacher Academy.

Barnett, D. W., Lentz Jr., F. E., & Macmann, G. (2000). Psychometric qualities of professional practice. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools: Theory, research, and clinical foundations* (2nd ed.) (pp. 355–386). New York: Guilford.

Beck, I. L., Omanson, R. C., & McKeown, M. G. (1982). An instructional redesign of reading lessons: Effects on comprehension. *Reading Research Quarterly, 17,* 462–481.

Bellinger, J. M., & DiPerna, J. C. (2011). Is fluency-based story retell a good indicator of reading comprehension? *Psychology in the Schools, 48,* 416–426.

Berninger, V., Abbott, R., Billingsley, F., & Nagy, W. (2001). Process underlying timing and fluency: Efficiency, automaticity, coordination, and morphological awareness. In M. Wolf (Ed.), *Dyslexia, fluency, and the brain* (pp. 388–414). Baltimore: York Press.

Berninger, V. W., Abbott, R. D., Vermeulen, K., & Fulton, C. M. (2006). Paths to reading comprehension in at-risk second-grade readers. *Journal of Learning Disabilities, 39*, 334–351.

Best, R. M., Floyd, R. G., & McNamara, D. S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading psychology, 29*, 137–164.

Bishop, A. G. (2003). Prediction of first-grade reading achievement: A comparison of fall and winter kindergarten screenings. *Learning Disability Quarterly, 26*, 189–202.

Blachowicz, C., & Ogle, D. (2008). *Reading comprehension: Strategies for independent learners* (2nd ed.). New York: Guilford.

Blachowicz, C., & Ogle, D. (2001). *Reading comprehension: Strategies for independent learners.* New York: Guilford.

Block, C. C. (2005). What are metacognitive assessments. In S. E. Israel, C. C. Block, K. L. Bauserman, & K. Kinnucan-Welsch (Eds.), *Metacognition in literacy learning: Theory, assessment, instruction, and professional development* (pp. 83–100). Mahwah, NJ: Lawrence Erlbaum Associates.

Buck, J., & Torgeson, J. (2003). *The relationship between performance on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test* (Technical Report 1). Tallahassee, FL: Florida Center for Reading Research.

Burke, M. D., & Hagan-Burke, S. (2007). Concurrent criterion-related validity of early literacy indicators for middle of first grade. *Assessment for Effective Intervention, 32*, 66–77.

Bursuck, W. D., & Damer, M. (2007). *Reading instruction for students who are at risk or have disabilities.* Boston: Pearson.

Cain, K., & Oakhill, J. (2006a). Assessment matters: Issues in the measurement of reading comprehension. *British Journal of Educational Psychology, 76*, 697–708.

Cain, K., & Oakhill, J. (2006b). Profiles of children with specific reading comprehension difficulties. *British Journal of Educational Psychology, 76*, 683–696.

Caldwell, J. S. (2008). *Comprehension assessment: A classroom guide.* New York: Guilford.

Caldwell, J. S., & Leslie, L. (2005). *Intervention strategies to follow informal reading inventory assessment: So what do I do now?* Boston: Pearson.

Campbell, J. R. (2005). Single instrument, multiple measures: Considering the use of multiple item formats to assess reading comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 347–368). Mahwah, NJ: Lawrence Erlbaum Associates.

Carlisle, J. F. (1993). Understanding passages in science textbooks: A comparison of students with and without learning disabilities. *National Reading Conference Yearbook, 42*, 235–242.

Cash, M. M., & Schumm, J. S. (2006). Making sense of knowledge: Comprehending expository text. In J. S. Schumm (Ed.), *Reading assessment and instruction for all learners* (pp. 262–296). New York: Guilford.

Castillo, J. M., Torgeson, J. K., Powell-Smith, K. A., & Al Otaiba, S. (2003). Relationships of five reading fluency measures to reading comprehension in first through third grade. Manuscript in preparation.

Chall, J. S. (1983). *Stages of Reading Development* (2nd ed.). Fort Worth, TX: Harcourt.

Christ, T. J., & Silberglitt, B. (2007). Estimates of the standard error of measurement for curriculum-based measures of oral reading fluency. *School Psychology Review, 36*, 130–146.

Cohen, E. J. K. (1996). *The effects of a holistic graphophonic intervention on the decoding performance of children with reading disabilities.* Unpublished doctoral dissertation, Florida International University. Quoted in Pazos-Rego, A. M. (2006). The alphabetic principle, phonics, and spelling: Teaching students the code. In J. S. Schumm (Ed.), *Reading assessment and instruction for all learners* (pp. 118–162). New York: Guilford.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 115–159.

Compton, D. L., Appleton, A. C., & Hosp, M. K. (2004). Exploring the relationship between text-leveling systems and reading accuracy and fluency in second-grade students who are average and poor decoders. *Learning Disabilities Research & Practice, 19*, 176–184.

Compton, D. L., Fuchs, D., Fuchs, L.S., Bouton, B., Gilbert, J. K., Barquero, L. A. et al. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal Educational Psychology, 102,* 327–340.

Copmann, K. S. P., & Griffith, P. L. (1994). Event and story structure recall by children with specific learning disabilities, language impairments, and normally achieving children. *Journal of Psycholinguistic Research, 23,* 231–248.

CTB/McGraw-Hill. (2002). *TerraNova.* Monterey, CA: CTB/McGraw-Hill.

Crawford, L., Tindal, G., & Stieber, S. (2001). Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment, 7,* 303–323.

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading, 10,* 277–299.

Daneman, M., & Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple-choice reading comprehension tests such as the SAT. *Journal of Experimental Psychology: General, 130,* 208–223.

Data Recognition Corporation (DRC). (2007). *Technical report for the Pennsylvania System of School Assessment, 2006 Reading and Mathematics Grades 4, 6, 7.* (available at http://www.portal.state.pa.us/portal/server.pt/gateway/PTARGS_0_0_51030_0_0_43/http;/pubcontent.state.pa.us/publishedcontent/publish/cop_hhs/pde/pde_community_content/dsf_migration/k12/assessment/content/resources/technical_analysis/technical_analysis_materials_for_the_pssa.htm?qid=69576400&rank=5)

Data Recognition Corporation (DRC). (2009). *Technical report for the Pennsylvania System of School Assessment, 2008 Reading and Mathematics Grades 2, 4, 5, 6, 7, 8 and 11.* (available at http://www.portal.state.pa.us/portal/server.pt/gateway/PTARGS_0_0_51030_0_0_43/http;/pubcontent.state.pa.us/publishedcontent/publish/cop_hhs/pde/pde_community_content/dsf_migration/k12/assessment/content/resources/technical_analysis/technical_analysis_materials_for_the_pssa.htm?qid=69576400&rank=5)

Data Recognition Corporation (DRC). (2010). *Technical report for the 2010 Pennsylvania System of School Assessment.* (available at http://www.portal.state.pa.us/portal/server.pt/ gateway/PTARGS_0_0_51030_0_0_43/http;/pubcontent.state.pa.us/publishedcontent/publish/cop_hhs/pde/pde_community_content/dsf_migration/k12/assessment/content/resource/technical_analysis/technical_analysis_materials_for_the_pssa.htm?qid=69576400&rank=5)

Data Recognition Corporation (DRC). (2011). *Technical report for the 2011 Pennsylvania System of School Assessment.* (available at http://www.portal.state.pa.us/portal/server.pt/gateway/PTARGS_0_0_51030_0_0_43/http;/pubcontent.state.pa.us/publishedcontent/publish/cop_hhs/pde/pde_community_content/dsf_migration/k12/assessment/content/resources/technical_analysis/technical_analysis_materials_for_the_pssa.htm?qid=69576400&rank=5)

Davey, B. (1988). The nature of response errors for good and poor readers when permitted to reinspect text during question-answering. *American Educational Research Journal, 25,* 399–414.

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

Deno, S. L. (1986). Formative evaluation of individual student programs: A new role for school psychologists. *School Psychology Review, 15,* 358–374.

Deno, S. L. (1989). Curriculum-based measurement and alternative special education services: A fundamental and direct relationship. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp.1–17). New York: Guilford.

Deno S. L., & Mirkin, P. K. (1977). *Data-based program modification: A manual*. Reston, VA: Council for Exceptional Children.

Denton, C. A., Fletcher, J. M., Anthony, J. L., & Francis, D. J. (2006). An evaluation of intensive intervention for students with persistent reading difficulties. *Journal of Learning Disabilities, 39,* 447–466.

Dewitz, P., & Dewitz, P. K. (2003). They can read the words, but they can't understand: Refining comprehension assessment. *The Reading Teacher, 56, 5,* 422–435.

Diakidoy, I. N., Stylianou, P., Karefillidou, C., & Papageorgiou, P. (2005). The relationship between listening and reading: Comprehension of different types of text at increasing grade levels. *Reading Psychology, 26,* 55–80.

Durkin, D. (1980). *Teaching young children to read* (3rd ed.). Boston: Allyn & Bacon.

Durlak, J. A. (1997). Primary prevention programs in schools. *Advances in Clinical Child Psychology, 19,* 283–318.

Duke, N. K., & Kays, J. (1998). "Can I say 'once upon a time'?": Kindergarten children developing knowledge of information book language. *Early Childhood Research Quarterly, 13,* 295–318.

Dynamic Measurement Group (DMG). (2011a). *DIBELS Next assessment manual.* Eugene, OR: Author.

Dynamic Measurement Group (DMG). (2011b). *DIBELS Next technical adequacy supplement.* Eugene, OR: Author.

Dynamic Measurement Group (DMG). (2008). *DIBELS 6th edition technical adequacy information* (Technical Report No. 6). Eugene, OR: Author.

Dynamic Measurement Group (DMG). (2007a). *DIBELS training institute: Essential workshop. Module 8 administration and scoring of DIBELS oral reading fluency (DORF). Module 9 administration and scoring of retell fluency (RTF).* Eugene, OR: Author.

Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012, February 13). Reader-text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology.* Advance online publication.

edHelper. (2009). *Grade 3 reading comprehensions.* Retrieved from http://www.edhelper.com/.

Ekwall, E. E., & Shanker, J. L. (1989). *Teaching reading in the elementary school.* Columbus, OH: Merrill Publishing Company.

Elliott, S. N., Huai, N., & Roach, A. T. (2007). Universal and early screening for educational difficulties. Current and future approaches. *Journal of School Psychology, 45*, 137–161.

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences.* Pacific Grove, CA: Brooks/Cole Publishing.

Farr, R. (1999). Putting it all together: Solving the reading assessment puzzle. In S. J. Barrentine (Ed.), *Reading assessment: Principles and practices for elementary teachers* (pp. 44–56). Newark, DE: International Reading Association.

Fewster, S., & MacMillan, P. D. (2002). School-based evidence for the validity of curriculum-based measurement of reading and writing. *Remedial and Special Education, 23,* 149– 156.

Fleisher, L. S., Jenkins, J. R., & Pany, D. (1979). Effects on poor readers' comprehension of training in rapid decoding. *Reading Research Quarterly, 15,* 30–48.

Foorman, B. R., Francis, D. J., Shaywitz, S. E., Shaywitz, B. A., & Fletcher, J. M. (1997). The case for early reading interventions. In B. A. Blachman (Ed.), *Foundations of reading acquisition and dyslexia: Implications for early intervention* (pp. 243–264). Hillsdale, NJ: Erlbaum.

Francis, D. J. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology*, *88*, 3–17.

Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology, 46*, 315–342.

Fritschmann, N. S., Shapiro, E. S., & Thomas, L. B. (2010). *Unpublished data*. Bethlehem, PA: Lehigh University.

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57,* 488–500.

Fuchs, L. S., & Deno, S. L. (1992). Effects of curriculum within curriculum-based measurement. *Exceptional Children, 58,* 232–243.

Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review, 21*, 45–58.

Fuchs, L. S., & Fuchs, D. (2002). Curriculum-based measurement: Describing competence, enhancing outcomes, evaluating treatment effects, and identifying treatment nonresponders. *Peabody Journal of Education*, *77*(2), 64–84.

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

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9(2),* 20–28.

Gajria, M., Jitendra, A. K., Sood, S., & Sacks, G. (2007). Improving comprehension of expository text in students with LD: A research synthesis. *Journal of Learning Disabilities, 40*, 210–225.

Gambrell, L. B., Koskinen, P. S., & Kapinus, B. A. (1991). Retelling and the reading comprehension of proficient and less-proficient readers. *Journal of Educational Research, 84,* 356–362.

Gambrell, L. B., Pfeiffer, W. R., & Wilson, R. M. (1985). The effects of retelling upon reading comprehension and recall of text information. *Journal of Educational Research, 78,* 216–220.

Gardill, M. C., & Jitendra, A. K. (1999). Advanced story map instruction: Effects on the reading comprehension of students with learning disabilities. *The Journal of Special Education, 33,* 2–17.

Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. (1982, 1983). *Stanford achievement test.* Iowa City: Harcourt, Brace, Jovanovich.

Gersten, R., Fuchs, L. S., Williams, J. P., & Baker, S. (2001). Teaching reading comprehension strategies to students with learning disabilities: A review of research. *Review of Educational Research, 71*, 279–320.

Glenn, C. G. (1980). Relationship between story content and structure. *Journal of Educational Psychology, 72,* 550–560.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45,* 117–135.

Good, R. H., & Kaminski, R. A. (Eds.). (2002). *Dynamic indicators of basic early literacy skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement.

Good, R. H., Simmons, D., & Kame'enui, E. (2001). The importance of decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies in Reading, 5,* 257–288.

Good, R. H., Simmons, D. C., Kame'enui, E. J., & Wallin, J. (2002). *Summary of decision rules for intensive, strategic, and benchmark instructional recommendations in kindergarten through third grade* (Technical Report No. 11). Eugene, OR: University of Oregon.

Gough, P. (1996). How children learn to read and why they fail. *Annals of Dyslexia, 46,* 3–20.

Graesser, A. C., McNamara, D. S., & Kuilkowich, J. M. (2011). Coh-metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40,* 223–234.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text? In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 82–98). New York: Guilford.

Gresham, F. M., & Witt, J. C. (1997). Utility of intelligence tests for treatment planning, classification, and placement decisions: Recent empirical findings and future directions. *School Psychology Quarterly*, *12*, 249–267.

Grimm, L. G., & Yarnold, P. R. (1995). *Reading and understanding multivariate statistics.* Washington, D.C.: American Psychological Association.

Hagtvet, B. E. (2003). Listening comprehension and reading comprehension in poor decoders: Evidence for the importance of syntactic and semantic skills as well as phonological skills. *Reading and Writing: An Interdisciplinary Journal, 16,* 505–539.

Hall, K. M., Sabey, B. L., & McClellan, M. (2005). Expository text comprehension: Helping primary-grade teachers use expository texts to full advantage. *Reading Psychology, 26*, 211–234.

Hamilton, C., & Shinn, M. R. (2003). Characteristics of word callers: An investigation of the accuracy of teachers' judgments of reading comprehension and oral reading skills. *School Psychology Review, 32*, 228–240.

Hansen, C. L. (1978). Story retelling used with average and learning disabled readers as a measure of reading comprehension. *Learning Disability Quarterly,1, 62–69.*

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children.* Baltimore: Paul H. Brookes.

Hidi, S., & Anderson, V. (1986). Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of Educational Research, 56,* 473–493.

Hintze, J. L. (2008). Logistic regression. In Author, *PASS help system* (860-1–860-16). Kaysville, UT: NCSS.

Hintze, J. M., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review, 34*, 372–386.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2*(2), 127–160.

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

Hsieh, F. Y., Bloch, D. A., & Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine, 17*, 1623–1634.

Irwin, J. W. (1979). Fifth grade readers' comprehension of explicit and implicit connective propostions. *Journal of Reading Behavior, 11*, 261–271.

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36*, 582–600.

Jenkins, J. R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Reading aloud and maze. *Exceptional Children, 59,* 421–432.

Johnson, P. H. (1983). *Reading comprehension assessment: A cognitive basis.* Newark, DE: International Reading Association.

Johnson, L., Graham, S., & Harris, K. R. (1997). The effects of goal setting and self-instruction on learning a reading comprehension strategy: A study of students with learning disabilities. *Journal of Learning Disabilities, 30,* 80–91.

Johnston, P. B. (1982). *Implications of basic research for the assessment of reading comprehension* (Technical Report No. 206). Urbana-Champaign: University of Illinois, Center for the Study of Reading. (ERIC ED 217402)

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80,* 437–447.

Juel, C. (1991). Beginning reading. In R. Barr, M. L. Kamil, P. B. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2) (pp. 759–788). White Plains, NY: Longman.

Kame’enui, E. J., Carnine, D. W., & Freschi, R. (1982). Effects of text construction and instructional procedures for testing word meanings on comprehension and recall. *Reading Research Quarterly, 17,* 367–388.

Katz, S., Blackburn, A. B., & Lautenschlager, G. J. (1991). Answering reading comprehension items without passages on the SAT when items are quasi-randomized. *Educational and Psychological Measurement, 52,* 747–754.

Katz, S., Lautenschlager, G. J., Blackburn, A. B., & Harris, F. H. (1990). Answering reading comprehension items without passages on the SAT. *Psychological Science, 1*, 122–127.

Kazdin, A. E. (2003). *Research design in clinical psychology.* (4th ed.). Boston, MA: Allyn and Bacon.

Keenan, J. M., & Betjemann, R. S. (2006). Comprehending the Gray Oral Reading Test without reading it: Why comprehension tests should not include passage-independent items. *Scientific Studies of Reading, 10*, 363–380.

Keller-Margulis, M. A., Shapiro, E. S., & Hintze, J. M. (2008). Long-term diagnostic accuracy of curriculum-based measures in reading and mathematics. *School Psychology Review, 37,* 374–390.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge, UK: Cambridge University Press.

Kintsch, W., & Kintsch, E. (2005). Comprehension. In S. G. Paris, & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 71–92) Mahwah, NJ: Lawrence Erlbaum Associates.

Klingner, J. K. (2004). Assessing reading comprehension. *Assessment for Effective Intervention*, *29*(4), 59–67.

Klingner, J. K., Vaughn, S., & Boardman, A. (2007). *Teaching reading comprehension to students with learning difficulties*. New York: Guilford.

Kranzler, J. H., Miller, M. D., & Jordan, L. (1999). An examination of racial/ethnic and gender bias on curriculum-based measurement of reading. *School Psychology Quarterly*, *14,* 327–342.

Kucer, S. B. (2011). Going beyond the author: What retellings tell us about comprehending narrative and expository texts. *Literacy*, *45,* 62–69.

Kutner, M., Greenberg, E., Jin, Y., Boyle, B., Hsu, Y., & Dunleavy, E. (2007). *Literacy in everyday life: Results from the 2003 National Assessment of Adult Literacy* (NCES 2007-480). U. S. Department of Education. Washington, DC: National Center for Education Statistics.

Licht, M. H. (1995). Multiple regression and correlation. In L. G. Grimm & P. R. Yarnold (Eds.). *Reading and understanding multivariate statistics* (pp. 19–64). Washington, DC: American Psychological Association.

MacGinitie, W. H., Kamons, J., Kowalski, R. L., MacGinitie, R. K., & McKay, T. (1978). *Gates-MacGinitie Reading Tests* (2nd ed.). Chicago: Riverside.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). *Gates-MacGinitie Reading Tests (GMRT).* Itasca, IL: Riverside Publishing.

Macy, M. G., Bricker, D. D., & Squires, J. K. (2005). Validity and reliability of a curriculum-based assessment approach to determine eligibility for part c services. *Journal of Early Intervention, 28*, 1–16.

Marcotte, A. M., & Hintze, J. M. (2009). Incremental and predictive utility of formative assessment methods of reading comprehension. *Journal of School Psychology*, *47*, 315–335.

Margolis, H. (2004). Struggling readers: What consultants need to know. *Journal of Educational and Psychological Consultation, 15*, 191–204.

Maria, K. (1990). *Reading comprehension instruction: Issues and strategies*. Parkton, MD: York Press.

Marr, M. B., & Gormley, K. (1982). Children's recall of familiar and unfamiliar text. *Reading Research Quarterly, 18,* 89–104.

Marston, D. B. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18–78). New York: Guilford Press.

McGill-Franzen, A. (1987). Failure to learn to read: Formulating a policy problem. *Reading Research Quarterly, 22,* 475–490.

McGlinchey, M. T., & Hixon, M. D., (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review, 33*, 193–203.

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

McKenna, M. K., & Good, R. H. (2003). *Assessing reading comprehension: The relation between DIBELS Oral Reading Fluency, DIBELS Retell Fluency, and Oregon State Assessment scores.* Eugene: University of Oregon.

Medina, A. L., & Pilonieta, P. (2006). Once upon a time: Comprehending narrative text. In J. S. Schumm (Ed.), *Reading assessment and instruction for all learners* (pp. 222–261). New York: Guilford.

Meisinger, E. B., Bradley, B. A., Schwanenflugel, P. J., Kuhn, M. R., & Morris, R. D. (2009). Myth and reality of the word caller: The relation between teacher nominations and prevalence among elementary school children. *School Psychology Quarterly, 24*, 147–159.

MetaMetrics Inc. (2008). *The Lexile Analyzer.* Retrieved from http://www.lexile.com/analyzer.

Meyer, B. J. F., Brandt, D. M., & Bluth, G. J. (1980). Use of top-level structure in text: Key for reading comprehension of ninth-grade students. *Reading Research Quarterly, 16,* 72–103.

Micro Power & Light Co. (2008). *The Spache formula version 1.3.* Dallas, TX: Author.

Morrow, L. M. (1985). Retelling stories: A strategy for improving young children's comprehension, concept of story structure, and oral language complexity. *The Elementary School Journal, 85*, 646–661.

Morrow, L. M., Sisco, L. J., & Smith, J. K. (1992). The effect of mediated story retelling on listening comprehension, story structure, and oral language development in children with learning disabilities. *National Reading Conference Yearbook, 41,* 435–443.

Moss, B. (1997). A qualitative assessment of first graders' retelling of expository text. *Reading Research and Instruction, 37,* 1–13.

National Reading Panel (NRP). (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction.* Washington, DC: National Institute of Child Health and Human Development. (available at http://www.nationalreadingpanel.org/ Publications/summary.htm)

Nation, K. (2005). Children's reading comprehension difficulties. In M. J. Snowling, & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 248–265). Malden, MA: Blackwell.

No Child Left Behind (NCLB) Act of 2002. Public Law 107-110.

Oakhill, J., & Cain, K. (2007). Introduction to comprehension development. In K. Cain & J. Oakhill (Eds.), *Children's comprehension problems in oral and written language: A cognitive perspective* (pp. 3–40). New York: Guilford.

Olson, M. W. (1985). Text type and reader ability: The effects on paraphrase and text-based inference questions. *Journal of Reading Behavior, 17,* 199–214.

Pallant, J. (2007). *SPSS survival manual* (3rd ed.). New York: McGraw-Hill.

Paris, A. H., & Paris, S. G. (2003). Assessing narrative comprehension in young children. *Reading Research Quarterly, 38,* 36–76.

Paris, S. G., Wasik, B., & Turner, J. C. (1996). The development of strategic readers. In R. Barr, M. L. Kamil, P. B. Mosenthal, & P. D. Pearson, *Handbook of reading research* (Vol. 2) (pp. 609-640). Mahwah, NJ: Lawrence Erlbaum Associates.

Pearson Education Inc. (2008). *AIMSweb. Reading-CBM.* Retrieved from http://www.aimsweb.com.

Pearson, P. D., & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices-past, present, and future. In S. G. Paris, & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 13–69). Mahwah, NJ: Lawrence Erlbaum Associates.

Peng, C. Y., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research, 96*, 3–14.

Pennsylvania State Board of Education. (2007). *Math, reading, science and writing performance level cut scores.* Retrieved from http://www.portal.state.pa.us/portal/server.pt/community/cut_scores/7441.

Pennsylvania State Board of Education. (2005). *Pennsylvania system of school assessment grade 3 reading performance level descriptors.* Retrieved from http://www.portal.state.pa.us/portal/server.pt?open=514&objID=507629&mode=2.

Powell-Smith, K. A., Good, R. H., Latimer, R. J., Dewey, E. N., & Kaminski, R. A. (2011). *DIBELS Next benchmark goals study* (Technical Report No. 11). Eugene, OR: Dynamic Measurement Group.

Prescott, G. A., Balow, I. H., Hogan, T. P., & Farr, R. C. (1984). *Metropolitan Achievement Tests.* San Antonio, TX: Psych Corp.

Pressley, M. (2000). What should comprehension instruction be the instruction of? In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research,* (Vol. 3) (pp. 545–561). Mahwah, NJ: Lawrence Erlbaum Associates.

Pressley, M., Hilden, K. R., & Shankland, R. (2005). *An evaluation of the grade-3 DIBELS oral fluency measure* (Technical Report). East Lansing MI: Michigan State University, College of Education, Literacy Achievement Research Center.

Rabren, K., Darch, C., & Eaves, R. C. (1999). The differential effects of two systematic reading comprehension approaches with students with learning disabilities. *Journal of Learning Disabilities, 32,* 36–47.

Rack, J. P., Snowling, M. J., & Olson, R. K. (1992). The nonword reading deficit in developmental dyslexia: A review. *Reading Research Quarterly, 27,* 28–53

RAND Reading Study Group (RRSG). (2002). *Reading for understanding: Toward an R&D program in reading comprehension.* Washington, DC: RAND Education.

Randi, J., Grigorenko, E. L., & Sternberg, R. J. (2005). Revisiting definitions of reading comprehension: Just what is reading comprehension anyway. In S. E. Israel, C. C. Block, K. L. Bauserman, & K. Kinnucan-Welsch (Eds.), *Metacognition in literacy learning: Theory, assessment, instruction, and professional development* (pp. 19–39). Mahwah, NJ: Lawrence Erlbaum Associates.

Rapp, D. N., van den Broek, P., McMaster, K. L., Panayiota, K., & Espin, C. A. (2007). Higher-order comprehension processes in struggling readers: A perspective for research and intervention. *Scientific Studies of Reading, 11,* 289–312.

Rasinski, T. V. (1990). Investigating measures of reading fluency. *Educational Research Quarterly, 14*(3), 37–44.

Rathvon, N. (2004). *Early reading assessment: A practitioner's handbook*. New York: Guilford.

Reed, D. K. (2011). A review of the psychometric properties of retell instruments. *Educational Assessment, 16*, 123–144.

Reed, D. K., & Petscher, Y. (in press). Which reading tests are useful to middle school teachers: Balancing concurrent and face validity. *Assessment for Effective Instruction.*

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47,* 427–469.

Riedel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly, 42*, 546–567.

Ringler, L. H., & Weber, C. (1984). A language-thinking approach to reading. New York: Harcourt Brace Jovanovich.

Risko, V. J., & Alvarez, M. C. (1986). An investigation of poor readers' use of a thematic strategy to comprehend text. *Reading Research Quarterly, 21,* 298–316.

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

Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgeson, J. K. (2008). Accuracy of DIBELS Oral Reading Fluency for predicting third grade reading comprehension outcomes. *Journal of School Psychology, 46,* 343–366.

Royer, J. M. (1990). The sentence verification technique: A new direction in the assessment of reading comprehension. In S. Legg & J. Algina (Eds.), *Cognitive assessment of language and math outcomes: Advances in discourse processes* (Vol. 36) (pp. 144–191). Westport, CT: Ablex Publishing.

Royer, J. M., & Lynch, D. J. (1983). The misuses and appropriate uses of norm-referenced tests of reading comprehension. *Reading Psychology, 3,* 131–142.

Scanlon, D. M., Vellutino, F. R., Small, S. G., Fanuele, D. P., & Sweeney, J. M. (2005). Severe reading difficulties – can they be prevented? A comparison of prevention and intervention approaches. *Exceptionality, 13,* 209–227.

Schisler, R., Joseph, L. M., Konrad, M., & Alber-Morgan, S. (2010). Comparison of the effectiveness and efficiency of oral and written retellings and passage review as strategies for comprehending text. *Psychology in the Schools, 47,* 135–152.

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology, 57,* 1–10.

Shanahan, T., Kamil, M. L., & Tobin, A. W. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly, 17,* 229–255.

Shankweiler, D., Crain, S., Katz, L., Fowler, A. E., Liberman, A. M., Brady, S. A., et al. (1995). Cognitive profiles of reading-disabled children: Comparison of language skills in phonology, morphology, and syntax. *Psychological Science, 6,* 149–156.

Shannon, P., Kame’enui, E. J., & Baumann, J. F. (1988). An investigation of children’s ability to comprehend character motives. *American Educational Research Journal, 25,* 441–462.

Shapiro, E. S. (2004). *Academic skills problems: Direct assessment and intervention* (3rd ed.). New York: Guilford.

Shapiro, E. S., Fritschmann, N. S., Thomas, L. B., Hughes, C. L., & McDougal, J. (2010). Preliminary development of a benchmarking measure for reading comprehension. Paper presented at the annual convention of the American Psychological Association, San Diego, CA.

Shapiro, E. S., Keller, M. A., Edwards, L., Lutz, G., & Hintze, J. M. (2006). General outcome measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment, 42*, 19–35.

Shapiro, E. S., Solari, E., & Petscher, Y. (2008). Use of a measure of reading comprehension to enhance prediction on the state high stakes assessment. *Learning and Individual Differences, 18*, 316–328.

Shaw, R., & Shaw, D. (2002). *DIBELS oral reading fluency-based indicators of third grade reading skills for Colorado State Assessment Program (CSAP).* (Technical Report). Eugene, OR: University of Oregon.

Shinn, M. R. (Ed.). (1989). *Curriculum-based measurement: Assessing special children.* New York: Guilford.

Shinn, M. R. (Ed.). (1998). *Advanced applications of curriculum-based measurement.* New York: Guilford.

Shinn, M. R. (2002). Best practices in using curriculum-based measurement in a problem-solving model. In A. Thomas, & J. Grimes (Eds.), *Best practices in* *school psychology IV* (Vol. 2) (671–698). Bethesda, MD: National Association of School Psychologists.

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

Shinn, M. M., & Shinn, M. R. (2002). *AIMSweb training workbook: Administration and scoring of reading curriculum-based measurement (R-CBM) for use in general outcome measurement.* (available at http://www.aimsweb.com/uploads/news/ id24/rcbm_summary _of_ validity.pdf)

Short, E. J., Yeates, K. O., & Feagans, L. V. (1992). The generalizability of story grammar training across setting and tasks. *Journal of Behavioral Education, 2,* 105–120.

Sibley, D., Biwer, D., & Hesch, A. (2001). *Unpublished data*. Arlington Heights, IL: Arlington Heights School District 25.

Silberglitt, B., Burns, M. K., Madyun, N. H., & Lail, K. E. (2006). Relationship of reading fluency assessment data with state accountability test scores: A longitudinal comparison of grade levels. *Psychology in the Schools, 43,* 527–535.

Simmons, D. C., Kuykendall, K., King, K., Cornachione, C., & Kame’enui, E. J. (2000). Implementation of a schoolwide reading improvement model: "No one ever told us it would be this hard!” *Learning Disabilities Research & Practice, 15*, 92–100.

Snow, C. E., & Sweet, A. P. (2003). Reading for comprehension. In A. P. Sweet, & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 1–11). New York: Guilford.

Snyder, L., Caccamise, D., & Wise, B. (2005). The assessment of reading comprehension: Considerations and cautions. *Topics in Language Disorders, 25,* 33–50.

Speece, D. L. (2005). Hitting the moving target known as reading development: Some thoughts on screening children for secondary interventions. *Journal of Learning Disabilities, 38,* 487–493.

Speece, D. L., Ritchey, K. D., Silverman, R., Schatschneider, C., Walker, C. Y., & Andrusik, K. N. (2010). Identifying children in middle childhood who are at risk for reading problems. *School Psychology Review, 39,* 258–276.

Spiro, R. J., & Taylor, B. M. (1980). *On investigating children's transition from narrative to expository discourse: The multidimensional nature of psychological text classification* (Technical Report No. 195). Urbana-Champaign: University of Illinois, Center for the Study of Reading. (ERIC ED 199666)

Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review, 30*, 407–419.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21,* 360–407.

Stanovich, K. E. (1992). Speculations on the causes and consequences of individual differences in early reading acquisition. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds). *Reading acquisition* (pp. 307–342). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stein, B. L., & Kirby, J. R. (1992). The effects of text absent and text present conditions on summarization and recall of text. *Journal of Reading Behavior, 24*, 217–232.

Stevens, R. J., Slavin, R. E., & Farnish, A. M. (1991). The effects of cooperative learning and direct instruction in reading comprehension strategies on main idea identification. *Journal of Educational Psychology, 83,* 8–16.

Success for All Foundation. (2007). *4Sight reading and math benchmarks 2006–2007 technical report for Pennsylvania.* Baltimore: Success for All Foundation.

Success for All Foundation. (2008). *4Sight reading and math benchmarks 2008–2009 technical report for Pennsylvania.* Baltimore: Success for All Foundation.

Sweet, A. P. (2005). Assessment of reading comprehension: The Rand Reading Study Group vision. In S. G. Paris, & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 3–12). Mahwah, NJ: Lawrence Erlbaum Associates.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson.

Tannenbaum, K. R., Torgesen, J. K., & Wagner, R. K. (2006). Relationships between word knowledge and reading comprehension in third-grade children. *Scientific Studies of Reading, 10,* 381–398.

Taylor, L. K., Alber, S. R., & Walker, D. W. (2002). The comparative effects of a modified self-questioning strategy and story mapping on the reading comprehension of elementary students with learning disabilities. *Journal of Behavioral Education, 11,* 69–87.

Tharp, R. G., & Gallimore, R. (1988). *Rousing minds to life: Teaching, learning, and schooling in social context.* Cambridge: Cambridge University Press.

Torgesen, J. K. (1998). Catch them before they fall: Identification and assessment to prevent reading failure in young children. *American Educator*, *22*(1-2), 32–39

Torgesen, J., Schirm, A., Castner, L., Vartivarian, S., Mansfield, W., Myers, D. et al. (2007). *National assessment of Title I, final report. Volume II: Closing the reading gap: Findings from a randomized trial of four reading interventions for striving readers.* Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Science, U.S. Department of Education.

Tunmer, W. E., & Hoover, W. A. (1992). Cognitive and linguistic factors in learning to read. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds). *Reading acquisition* (pp. 175–214). Hillsdale, NJ: Lawrence Erlbaum Associates.

University of Oregon Center on Teaching and Learning (2010). *Comprehension instruction: Sequencing comprehension skills.* Retrieved from http://reading.uoregon.edu/big_ ideas /comp/comp_sequence.php.

Vander Meer, C. D., Lentz, F. E., & Stollar, S. (2005). *The relationship between oral reading fluency and statewide achievement testing in Ohio* (Technical Report). Eugene, OR: University of Oregon.

Vanderstaay, S. L. (2006). Learning from longitudinal research in criminology and the health sciences. *Reading Research Quarterly, 41,* 328–350.

Vellutino, F. R. (2003). Individual differences as sources of variability in reading comprehension in elementary school children. In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 51–81). New York: Guilford.

Vosniadou, S., & Schommer, M. (1988). Explanatory analogies can help children acquire information from expository text. *Journal of Educational Psychology, 80*, 524–536.

Warren, L., & Fitzgerald, J. (1997). Helping parents to read expository literature to their children: Promoting main-idea and detail understanding. *Reading Research and Instruction*, *36*, 341–360.

Watson, T. S., & Skinner, C. H. (Eds.). (2004). *Encyclopedia of school psychology.* New York: Kluwer Academic/Plenum.

Whitehurst, G. J., & Massetti, G. M. (2004). How well does head start prepare children to learn to read? In E. Zigler & S. Styfco (Eds.), *The Head Start debates.* New Haven, CT: Yale University Press.

Wiederholt, J. L., & Bryant, B. R. (1992). *Gray Oral Reading Test* (3rd ed.). Austin, TX: PRO-ED.

Wiederholt, J. L., & Bryant, B. R. (2001). *Gray Oral Reading Test* (4th ed.). Austin, TX: PRO-ED.

Wiley, H. I., & Deno, S. L. (2005). Oral reading and maze measures as predictors of success for English language learners on state standards assessment. *Remedial and Special Education, 26*(4), 207–214.

Williams, J. L., Skinner, C. H., Floyd, R. G., Hale, A. D., Neddenriep, C., & Kirk, E. P. (2011). Words correct per minute: The variance in standardized reading scores accounted for by reading speed. *Psychology in the Schools, 48,* 87–101.

Williams, J. P. (2005). Instruction in reading comprehension for primary-grade students: A focus on text structure. *The Journal of Special Education, 39*(1), 6–18.

Williams, K. T. (2001). *Group Reading Assessment and Diagnostic Evaluation (GRADE).* Circle Pines, MN: American Guidance Services.

Wilson, J. (2005). *The relationship of Dynamic Indicators of Basic Early Literacy Skills (DIBELS) oral reading fluency to performance on Arizona Instrument to Measure Standards (AIMS).* (Research Brief). Tempe, AZ: Assessment and Evaluation Department, Tempe School District No. 3.

Winograd, P. N., Wixson, K. K., & Lipson, M. Y. (1989). *Improving basal reading instruction.* New York: Teachers College Press.

Wolfe, M. B. W., & Woodwyk, J. M. (2010). Processing and memory of information presented in narrative and expository texts. *British Journal of Educational Psychology, 80*, 341–362.

Wood, D. E. (2006). Modeling the relationship between oral reading fluency and performance on a statewide reading test. *Educational Assessment, 11,* 85–104.

Woodcock, R. M. (1987). *Woodcock Reading Mastery Test–Revised.* Circle Pines, MN: American Guidance Corp.

Zabrucky, K., & Ratner, H. H. (1992). Effects of passage type on comprehension monitoring and recall in good and poor readers. *Journal of Reading Behavior, 24,* 373–391.

Appendix A

Table A1

Summary of Research Using Free Oral Retell to Assess Reading Comprehension

Authors (Date)	Participants		Measures	Free Oral Retell Procedures/Scoring	Outcomes
	n	Grade			
Beck et al. (1982)	▪ 24 skilled readers ▪ 24 less-skilled readers ▪ conditions: (1) revised story or (2) original story	3rd	▪ 2 narrative texts with pictures ▪ free oral retell ▪ 35 forced-choice questions for each text: central, noncentral, & implied content	▪ participants read text silently, examiner assisted with any unfamiliar words ▪ asked to recall as much as they could of the story ▪ *text analysis – proportion of content recalled: gist of each central & noncentral content unit included in retell*	▪ recall was greater for central content vs. noncentral content ▪ question comprehension was greater for explicit vs. implied ▪ skilled readers performed better than less-skilled on both retell & forced-choice questions ▪ revised story group recalled more of the story & correctly answered more forced-choice questions
Best et al. (2008)	▪ 61 students	3rd	▪ 1 narrative text ▪ 1 expository text ▪ free oral retell ▪ cued oral retell ▪ 12 multiple-choice questions: literal & inferential ▪ Woodcock-Johnson Third Edition Tests of Achievement (WJ III ACH)	▪ participants read text silently within a 5-min period ▪ text was removed from view ▪ asked to report what they remembered about the passage they had just read & to give details like they were telling a friend ▪ recorded & transcribed ▪ 1–2 scorers per retell ▪ *text analysis – proportion of correct propositions/idea units included in retell*	▪ interscorer reliability: free & cued oral retell ≥ 90% ▪ comprehension was better (i.e., higher scores) for narrative text than expository text across different methods of assessment ▪ narrative text comprehension was more influenced by decoding skills ▪ expository text comprehension was more influenced by world knowledge ▪ children with high world knowledge also had high decoding skills
Burke & Hagan-Burke (2007)	▪ 213 students	1st	▪ 3 narrative passages ▪ Dynamic Indicators of Basic Early Literacy Skills (DIBELS): Phoneme Segmentation	▪ participants read text aloud ▪ text was removed from view ▪ asked to retell story in their own words ▪ *total # of words retold in 1 min*	▪ RTF correlated moderately with DORF (*r* = .69), SWE (*r* = .67), PDE (*r* = .59), & NWF (*r* = .54) ▪ RTF had low correlations with WUF (*r* = .31) & PSF (*r* = .26) ▪ RTF did not explain any variance in PDE or SWE after controlling for PSF, NWF, & DORF

			Fluency (PSF), Nonsense Word Fluency (NWF), DIBELS Oral Reading Fluency (DORF), Retell Fluency (RTF), & Word Use Fluency (WUF) ▪ Test of Word Reading Efficiency (TOWRE): Phonetic Decoding Efficiency (PDE) & Sight Word Efficiency (SWE)		▪ DORF & NWF had the strongest relationship with PDE & SWE ▪ DORF was single strongest predictor of PDE & SWE
Fleisher et al. (1979) *Study 2*	▪ 11 good readers ▪ 33 poor readers ▪ conditions: (1) poor readers with single word training, (2) poor readers with phrase training, (3) poor reader controls, or (4) good reader controls	4th & 5th	▪ 2 narrative texts ▪ free oral retell ▪ 6 inferential questions ▪ 6 factual questions ▪ cloze ▪ reading rate: words read per minute in isolation & context	▪ participants read text aloud, examiner corrected errors ▪ asked to tell everything they could remember about the story ▪ recorded & transcribed ▪ 2–3 scorers per retell ▪ *text analysis – total # of propositions/idea units included in retell*	▪ interscorer reliability: free oral retell mean = 91.8% ▪ good reader controls included significantly more propositions in their retell than poor readers with phrase training, poor readers with single word training, & poor reader controls ▪ decoding training (isolated words or phrases) significantly increased the decoding speed of single words ▪ decoding training did not improve comprehension – poor readers with either single word or phrase training performed no better on comprehension measures than did poor readers without training
Fuchs et al. (1988)	▪ 50 males with LD ▪ 16 males with emotional disturbance ▪ 4 males with mental retardation	4th–8th	▪ 4 narrative texts ▪ free oral & written retell ▪ ORF (i.e., words read correct) ▪ oral & written cloze ▪ 10 questions ▪ Stanford	▪ participants read text aloud within a 5-min period ▪ asked to tell in their own words what happened in the text within a 10-min period ▪ recorded & transcribed ▪ 1–2 scorers per retell ▪ *total # of words retold* ▪ *text analysis – % of content*	▪ interscorer reliability: total # of words retold for oral retell = 90%, % of content words retold for oral retell = 86%, and % of propositions included in oral retell = 89% ▪ alternative methods of scoring related comparably to SAT-7 Reading Comprehension – correlated highly with written retell & moderately with oral retell

			Achievement Test (SAT-7): Word Study Skills & Reading Comprehension	*words retold: exact match or synonym proper nouns, common nouns, verbs, adjectives, adverbs* ▪ *text analysis – % of propositions/idea units included in retell*	▪ counting total # of words retold was the most feasible scoring method, followed by % of content words retold ▪ ORF had highest correlations with SAT-7 Reading Comprehension & Word Study ▪ ORF had moderate to high correlations with the measures of reading comprehension ▪ ORF psychometrically useful method for monitoring overall reading growth
Gambrell et al. (1991)	▪ 24 proficient readers ▪ 24 less-proficient readers	4th	▪ 4 narrative stories at 2nd grade-level ▪ 4 narrative stories at 4th grade-level ▪ free oral retell ▪ cued recall: 4 explicit & 4 implicit comprehension questions	▪ participants read text silently ▪ 2-min period to think about how will tell story ▪ asked to retell into recorder so that younger children could listen to them tell the story ▪ recorded & transcribed ▪ 1–2 scorers per retell ▪ *text analysis – total # of propositions/idea units included in retell proper nouns counted once, noun referents & repetition were not counted* ▪ *text analysis – total # of positive elaborations & negative intrusions* ▪ *story elements – proportion of story structure elements recalled: setting, theme, plot, & resolution*	▪ interscorer reliability: free recall propositions = 94%, story structure elements = 95%, and cued recall questions = 92% ▪ 4 retelling practice sessions resulted in a significant increases in # of propositions recalled, proportion of story structure elements recalled, and # of cued-recall question answered correctly for both groups ▪ 4 retelling practice sessions resulted in significant improvements in the quantity & quality of the retelling of both groups ▪ proficient readers incorporated significantly more positive elaborations in session 4 than in session 1, whereas there was no change in positive elaborations for less-proficient readers ▪ practice in retelling generalized to different texts
Gambrell et al. (1985)	▪ 93 students ▪ conditions: (1) retelling or (2) illustrating	4th	▪ 5 expository passages ▪ immediate & 2-day delayed free oral retell to peer or free illustration retell ▪ cued oral retell: 10 literal & 10	▪ participants read text silently ▪ independently filled-in important idea & supporting details outline ▪ asked to retell/illustrate all of the important ideas from the story ▪ recorded & transcribed ▪ *text analysis – total # of*	▪ immediate & delayed oral retell group recalled significantly more agent/action, modifier, where/how/when, and proposed action than immediate & delayed illustration retell group ▪ immediate & delayed oral retell group remembered more and had a more complete & elaborate recall than immediate & delayed illustration retell

			inferential comprehension questions ▪ cued written illustration retell ▪ important idea & supporting details outline	*propositional categories recalled: agent/ action, modifier, where/ how/when, belongs to, conjoining, & proposed action*	group ▪ immediate illustration retell group recalled more agent/action, modifier, where/how/when, and proposed action than delayed illustration retell group ▪ 4 retelling practice sessions resulted in greater recall for oral retelling group than illustration retell group ▪ cued oral retelling group did better than cued illustration retell group on literal & inferential questions ▪ practice in retelling generalized to different texts
Hagtvet (2003) Study was conducted in Norway	▪ 24 good decoders ▪ 24 average decoders ▪ 24 poor decoders ▪ longitudinal followed from 4 to 9 years ▪ conditions: (1) listening or (2) reading	2nd (age 9 years)	▪ 1 orally presented story on tape recorder (listening) ▪ 1 written presented story (reading) ▪ free oral retell ▪ cloze ▪ all materials were in Norweigan ▪ phonemic awareness & complex syntax tests ▪ Wechsler Intelligence Scale for Children – Revised (WISC-R) Norweigan standardization: vocabulary, digit span, & prorated IQ	▪ participant either listened to story on a tape recorder or read the story ▪ asked to retell the story as completely as possible into recorder, told examiner would transcribe participant's story ▪ recorded & transcribed ▪ *story elements – max 3 points introduction, max 3 points cause/motive, max 2 points event 1, max 2 points event 2, max 1 point event 3, max 2 points event 4, max 2 points result, max 2 points ending 1, & max 3 points ending 2*	▪ poor decoders scored lower than average & good decoders on both listening & reading story retell and cloze tasks ▪ story retelling task: poor, average, & good decoders all performed slightly better on the listening version ▪ cloze task: poor, average, & good decoders all performed slightly better on the reading version ▪ overall good decoders had the highest mean scores on all measures, average decoders had slightly lower scores than good decoders, and poor decoders scored lower than average & good decoders on all measures ▪ significant moderate to high correlations amongst all measures ▪ vocabulary score was a significant predictor of reading story retell score ▪ syntax & phonemic awareness scores were significant predictors of performance on listening & reading cloze
Irwin (1979)	▪ 64 students ▪ conditions: (1) explicit causality connectives, (2) implicit causality	5th	▪ each group read 3 historical passages that included: explicit or implicit causality or time	▪ participants read text silently ▪ asked to recall passage ▪ recorded & transcribed ▪ 2 scorers ▪ *text analysis – total # of*	▪ interscorer reliability: # of connect propositions recalled ranged from .93 to 1.0 ▪ results provide no support for the notion that sentence length is related to

	connectives, (3) explicit time sequence connectives or (4) implicit time-sequence connectives		sequence connectives ▪ free oral retell ▪ forced-choice task	*critical propositions for the original explicitly included in retell*	comprehensibility or that implicit connectives are more difficult to comprehend that explicit ones ▪ participants had greater comprehension for the time-sequence relationship when they were stated explicitly or implicitly ▪ participants did not generally comprehend the causal relationships when they were stated explicitly or implicitly
Kame'enui et al. (1982)	▪ 60 students ▪ conditions: (1) easy vocab text with no training, (2) difficult vocab text with no training, (3) difficult vocab & redundant information text with no training, (4) difficult vocab text with vocab training, or (5) difficult vocab text with vocab & passage integration training	4th, 5th, & 6th	▪ 1 easy vocabulary passage ▪ 1 difficult vocabulary passage ▪ 1 difficult vocabulary & redundant information passage ▪ free oral retell + prompt for not responding ▪ vocabulary words & meanings on cards ▪ 5 literal & 4 inferential multiple-choice questions	▪ participants read text aloud, experimenter corrected decoding errors ▪ text was removed from view ▪ asked to retell everything that they remembered about the story in their own words ▪ recorded & transcribed ▪ 2 scorers ▪ *text analysis – total bits of information included in recall: difficulty vocabulary (8 bits), literal content (5 bits), & total retell (18 bits)*	▪ interscorer reliability: retell ranged from 87–99% ▪ substituting easy vocabulary words for difficult vocabulary words resulted in higher comprehension scores ▪ redundant information may improve comprehension for passages with difficult vocabulary words – experiment 1 difficult passage & redundant information group out performed difficulty passage only group on the multiple-choice test only; experiment 2 difficult passage & redundant information group out performed difficulty passage only group for both the multiple-choice test and retell bits of information ▪ instruction on difficult vocabulary improved comprehension for passages with difficult vocabulary words – participants who received passage integration training scored higher than those who read the difficult passage and did not receive training for the multiple-choice questions, recall of difficult vocabulary, & recall of total bits of information
Marcotte & Hintze (2009)	▪ 111 students	4th	▪ 3 narrative passages ▪ ORF ▪ DIBELS RTF ▪ sentence verification	▪ participants read text aloud ▪ text was removed from view ▪ asked to retell story in their own words ▪ *total # of words retold in 1 min*	▪ interscorer reliability: RTF only 33% were within 2-points of one another & only 46% were in 3-points of one another; range 0–15 points between raters; intraclass correlations for agreement of examiner's scores was.59 for RTF

			technique (SVT) ▪ written retell (WRT) ▪ maze/cloze (MZ) ▪ Group Reading Assessment and Diagnostic Evaluation (GRADE): Sentence Comprehension, Passage Comprehension, & Vocabulary ▪ Massachusetts Comprehensive Assessment System (MCAS) Language Arts		▪ RTF exhibited the weakest relationships with the other measures; WRT (*r* = .45), GRADE (*r* = .46), SVT & MZ (*r* = .47), & ORF (*r* = .49) ▪ RTF alone and in combination with the other measures did not contribute to explaining variance in the GRADE above and beyond the other measures ▪ ORF alone and in combination with the other comprehension measures predicted a significant proportion of the variance in GRADE (45–70%) ▪ MZ, SVT, & WRT may be indicators of students who have mastered of decoding skills but struggle with reading comprehension
Marr & Gormley (1982)	▪ 11 good readers ▪ 14 average readers ▪ 8 poor readers	4th	▪ 3 familiar topic & 3 unfamiliar topic expository passages ▪ oral retell ▪ oral reading accuracy level ▪ literal questions (pre-reading & post-reading/ probing)	▪ participants read text aloud ▪ asked to retell passage ▪ recorded & transcribed ▪ 2–3 scorers ▪ *text analysis – total # of propositions/idea units included in retell* ▪ *text analysis – total # of textual & scriptal propositions/idea units included in retell*	▪ retelling elicited text-based responses ▪ probing encouraged more responses based on prior knowledge ▪ comprehension ability & prior knowledge predicted comprehension performance ▪ prior knowledge of the topic was the strongest predictor of student's ability to make inferences & elaborations
Pressley et al. (2005)	▪ 191 students ▪ conditions: (1) standard DIBELS directions, (2) speed-emphasis directions, or (3) comprehension-emphasis directions	3rd	▪ 3 narrative texts ▪ DIBELS ORF & RTF	▪ participants read text aloud ▪ text was removed from view ▪ asked to retell story in their own words ▪ *total # of words retold in 1 min* ▪ transcribed retell & re-scored *total # of words retold in 1 min* ▪ *text analysis –total # of propositions/idea units*	▪ no significant group differences for ORF standard (mean = 113) vs. speed-emphasis (mean = 106) or speed-emphasis (mean = 106) vs. comprehension-emphasis (mean = 100) ▪ significant differences for ORF standard (mean = 113) vs. comprehension-emphasis (mean = 100) ▪ significant difference for ORF risk classification for standard vs. speed-emphasis and comprehension-emphasis,

				recalled	with standard scoring having a greater trend towards less risk ▪ ORF accounted for 20% of variance in TerraNova scores ▪ examinees appear to view ORF as a speed test ▪ significant difference between live RTF score & transcribed re-scoring (mean difference of 11 words); large effect size (mean = .95; range = .89 to 1.00) ▪ no significant group differences for total # of words retold across conditions ▪ total # of idea units include in retell was low for all stories & conditions ▪ RTF scores provided no predictive value relative to TerraNova performance
Rabren et al. (1999)	▪ 40 students with learning disabilities ▪ conditions: (1) explicit (direct instruction intervention) or (2) basal	4th	▪ 9 Aesop fables ▪ 3 modern fables ▪ 3 forms of text type: (1) textually explicit: character motive explicitly stated, (2) textually implicit: motive is implied, or (3) scriptually implicit: motive is neither explicit or implicit, drawn from prior knowledge ▪ daily story retells ▪ unit tests ▪ transfer, maintenance, & satisfaction measures	▪ participants read text aloud ▪ asked to retell passage ▪ experimenter took notes ▪ recorded & transcribed ▪ *story elements – scores for character motive scale 0–2: 0 = no information on character motive; 1 = partial information on character motive; 2 = complete & accurate information on character* motive ▪ *story elements – qualitative retelling profile indicated extent (none, low, moderate, or high degree) included or provided evidence of character motive*	▪ interscorer agreement = .87 character motive measure & .84 for qualitative retelling profile ▪ significant main effect for treatment group on both quantitative & qualitative retell measures; explicit group outperformed basal group for ability to retell character motivates across all 3 text types ▪ significant main effect for text type on both quantitative & qualitative retell measures; students performed better on textually explicit text type than textually implicit text type ▪ significant main effect for treatment group & text type on unit tests; explicit group outperformed basal group on unit test across all 3 text types & students performed better on textually explicit text type than textually implicit text type ▪ no significant main effect for treatment group on transfer & maintenance tests; however, there was a significant main effect for text type on transfer test –

					students performed better on textually explicit text type than textually implicit text type ▪ explicit rule-based instruction was effective for students with learning disabilities even when reading texts with various structures
Riedel (2007)	▪ 1,518 students	1st	▪ 3 narrative texts ▪ DIBELS Letter Naming Fluency (LNF), PSF, NWF, ORF, & RTF ▪ GRADE: vocabulary, comprehension, & oral-language skills	▪ participants read text aloud ▪ text was removed from view ▪ asked to retell story in their own words ▪ *total # of words retold in 1 min*	▪ beginning-of-year NWF & LNF slightly better predictors of GRADE comprehension at the end of 1st grade & TerraNova comprehension at the end of 2nd grade than PSF ▪ middle- and end-of-year ORF best predictor of GRADE comprehension at the end of 1st grade & TerraNova comprehension at the end of 2nd grade ▪ middle-and end-of year PSF poor predictor of GRADE comprehension at the end of 1st grade & TerraNova comprehension at the end of 2nd grade ▪ middle-and end-of year ORF had highest correlation of all the DIBELS measures with GRADE & TerraNova ▪ RTF was a weaker predictor of comprehension than ORF ▪ combination of LNF, PSF, NWF, & RTF measures with ORF did not substantially improve the predictive accuracy produced by ORF alone ▪ lack of empirical evidence for usefulness of RTF
Risko & Alvarez (1986) *Study 1*	▪ 86 below average readers ▪ conditions: A = thematic overview & guided instruction statements; B = thematic overview; C = guided	5th	▪ 1 expository passage from a social studies text ▪ immediate & 2-day delay free oral retell ▪ comprehension questions ▪ thematic overview	▪ participants read passage ▪ 4 min buffer task ▪ asked to tell everything they could remember about what they just read ▪ recorded & transcribed ▪ 3–5 raters ▪ *text analysis – textually explicit & textually implicit*	▪ interscorer reliability: textually explicit idea units = .98; textually implicit idea units = .96; textually implicit characteristics = .93 ▪ total # of textually explicit idea units included in retell yielded no significant main effects for treatment group or trial ▪ total # of textually implicit idea units included in retell yielded a significant

	instruction statements; D = passage only		▪ guided instruction statements	*propositions/idea units included in retell (1) rubric 0–3 textually explicit: 0 = incorrect response or no text-related information; 1 = vague paraphrase or small fragment of original unit; 2 = verbatim recall or good paraphrase of major part of original unit; 3 = verbatim recall of original unit; (2) textually implicit characteristics: attributes, goal statements, or causal & conditional relationships*	main effect for treatment group & trial – Groups A & B (both received thematic overview) outperformed Groups C & D, with Group A recalling more textually implicit idea units across both trials ▪ across Groups A, B, & C (treatment groups) majority of student responses were descriptions of attributes or goal statements; whereas Group D (control) the majority of student responses were descriptive ▪ Groups A, B, & C (treatment groups) included more causal & conditional relationships about thematic concepts compared to Group D (control) ▪ use of thematic organizer increased recall of text ideas and ability to elaborate upon implied information
Risko & Alvarez (1986) *Study 2*	▪ 24 below average readers ▪ conditions: A = thematic organizer; B = prereading questions	4th, 5th, & 6th	▪ 6 expository passages from 4 social studies texts ▪ thematic organizer ▪ prereading questions ▪ immediate & 2-day delay free oral retell ▪ 5 explicit & 5 implicit comprehension questions	▪ participants read passage, received 20 min for reading the passage and completing either organizer or prereading questions ▪ asked to tell everything they could remember about what they just read ▪ recorded & transcribed ▪ 3–5 raters ▪ *text analysis – total # of textually explicit & textually implicit propositions/idea units included in retell: (1) rubric 0–3 textually explicit: 0 = incorrect response or no text-related information; 1 = vague paraphrase or small fragment of original unit; 2 = verbatim recall or good paraphrase of major part of original unit; 3 = verbatim*	▪ interscorer reliability: retelling = .94 ▪ across the 5 passages used during the intervention phase, Group A (thematic organizer) outperformed Group B (prereading questions) on both the explicit & implicit comprehension questions ▪ for passage 6, the total # of textually explicit & implicit idea units included in retells yielded a significant main effect for treatment group & trial – Group A (thematic organizer) outperformed Group B (prereading questions) & Decline in both textually explicit & implicit idea units included in retell for both groups from Trial 1 to 2 ▪ for passage 6, total # of explicit & implicit comprehension questions answered correctly yielded a significant main effect for treatment group – overall Group A (thematic organizer) answered more explicit & implicit comprehension questions correctly; for explicit questions,

				recall of original unit; (2) textually implicit characteristics: attributes, goal statements, or causal & conditional relationships	Group A’s performance increased from Trial 1 to 2, whereas Group B’s performance decreased from Trial 1 to 2; for implicit questions, Group A’s performance decreased from Trial 1 to 2, whereas Group B’s performance increased from Trial 1 to 2 ▪ use of thematic organizer increased recall of text ideas, ability to elaborate upon implied information, and performance on literal & inferential comprehension questions
Roberts et al. (2005)	▪ 86 students	1st	▪ 2 narrative passages ▪ Vital Indicators of Progress (VIP) ORF (wcpm) ▪ VIP RTF ▪ Woodcock Diagnostic Reading Battery (WDRB): letter-word identification, word attack, & passage comprehension	▪ participants read text aloud ▪ text was removed from view ▪ asked to retell story in their own words ▪ *total # of words retold in 1 min*	▪ alternate-form reliability of RTF = .57 ▪ ORF correlated with WDRB Broad Reading Cluster .75 & .72 ▪ RTF correlated with WDRB Broad Reading Cluster .47 & .43 ▪ average RTF correlated with average ORF .61 ▪ ORF alone explained about 57% of variance in WDRB Broad Reading Cluster standard scores ▪ Adding RTF to ORF explained an additional 1% of the variance in WDRB Broad Reading Cluster standard scores = 58% for combination vs. 57% for ORF alone ▪ RTF as comprehension check for ORF ▪ ORF displayed a stronger relation with WDRB broad Reading Cluster for students with consistent retell (explained 65% of variance), than for students with inconsistent retell (explained 17% of variance); however, several potential confounds that may have impacted this finding are noted
Schisler et al. (2010)	▪ 5 general education students ▪ conditions: (1) repeated reading	3rd	▪ 45 reading passages ▪ oral retell ▪ written retell	▪ participants read text aloud ▪ repeated reading with drill-error correction procedure for mispronunciations or	▪ no significant difference across the three conditions for the total amount of time it took to complete the activities – it took participants a similar amount of time to

	with passage review, (2) repeated reading with oral retell, or (3) repeated reading with written retell		▪ passage review ▪ ORF ▪ multiple-choice comprehension questions: 5 literal & 5 inferential	omissions ▪ text was removed from view ▪ asked to tell about what just read within 3-min period ▪ *total number of minutes spent retelling* ▪ *total number of seconds it took to complete condition from when student began initial reading of the passage until student finished oral retell* ▪ *rate: # of comprehension questions answered correctly/time spent retelling*	read, reread passage, and use strategy across 3 conditions & participants had a similar amount of reading errors across the 3 conditions ▪ participants answered more comprehension questions correctly per minute of instructional time with the oral retelling condition than with the written retelling & passage review conditions ▪ oral retelling condition took the least amount of instructional time to implement, with most participants completing their oral retells within the 3-min period ▪ participants answered more literal questions than inferential questions correctly per minute of instructional time ▪ participants answered more literal & inferential questions correctly per minute of instruction time with the oral retelling strategy, followed closely by written retelling, and the least amount with the passage review condition ▪ both participants & teachers showed preference for retelling over passage review ▪ oral & written retells were found to be an effective strategy for improving comprehension of text
Shannon et al. (1988)	▪ 45 average or above average readers	2nd, 4th, & 6th	▪ 9 fables at each grade level (reading or listening) ▪ main character's motivation was either: (1) textually explicit, (2) textually implicit, or (3) scriptally implicit ▪ free oral retell ▪ detail-cue &	▪ participants read or listened to the text ▪ asked to retell as much about text as they could remember ▪ recorded & transcribed ▪ 2 scorers ▪ *story elements – identification of main character's motivation scored on 3-point scale: 0 = retell does not include explicit reference main character's*	▪ interscorer reliability: retell = 80% ▪ text type affected students comprehension ▪ students retold more character motives after reading textually explicit fables than after reading textually implicit fables or scriptally implicit fables ▪ students retold more detail-cues that were necessary to understand the main character's motivation after reading scriptally implicit fables and textually implicit fables as compared to textually

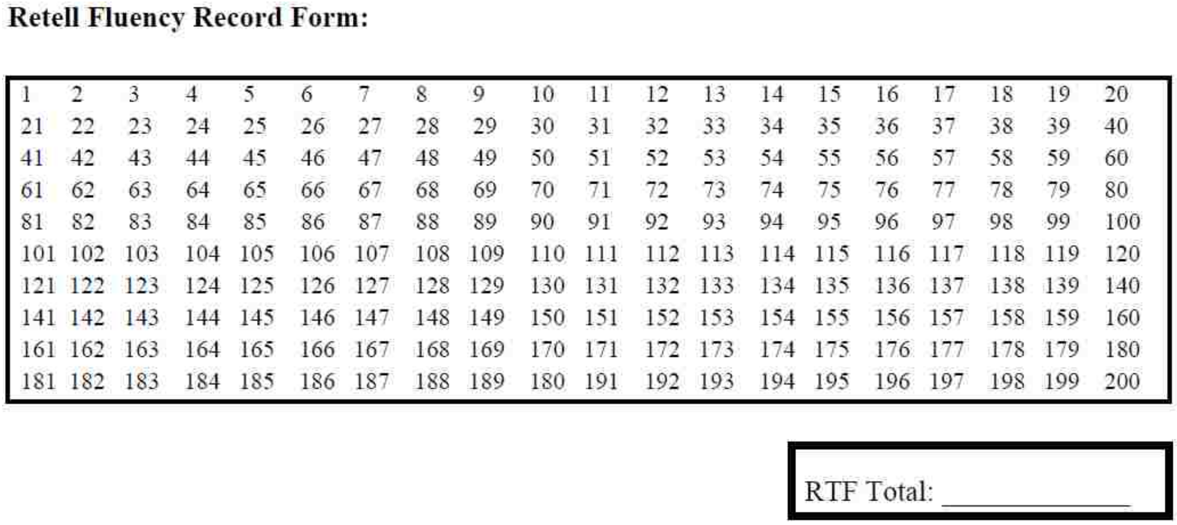
			motive comprehension questions	*motivation, 1 = partial reference, or 2 = complete reference* ▪ *story elements– identification of detail-cues (actions, events, descriptions) that were necessary to understand main character's motivation scored on 5-point scale: 0 = no detail-cues, 0.5 = one detail-cue, 1 = two detail-cues, 1.5 = three detail-cues, or 2 =four detail-cues*	explicit fables ▪ students answered more character motive questions correctly after reading textually explicit fables than they did after reading either textually implicit fables or scriptally implicit fables ▪ grade level affected students' abilities to answers comprehension questions about the main character's motivation ▪ students in Grade 6 outperformed students in Grade 2 & 4 on the character motive questions ▪ mode of presentation (reading vs. listening) did not affect students' ability to recall or answer comprehension questions about character's motivation or important detail-cues
Short et al. (1992)	▪ 36 students ▪ conditions: (1) experimental (story grammar training) or (2) control (no training)	4th & 5th	▪ 6 narrative passages: 5 canonical sequence & 1 flashback ▪ 3 expository passages ▪ 1 generalization passage not simple narrative or expository passage ▪ delayed free oral recall ▪ pre & post-training knowledge of story components ▪ note taking during story reading to assist with recall ▪ WISC-R information subtest ▪ delayed 14 short-answer questions	▪ participants read text aloud ▪ study passage until felt prepared for recall ▪ 7-minute delay WISC-R information subtest administered ▪ asked to retell text ▪ *story elements – 4-point scale for inclusion of story components: setting & main characters, initiating event, internal response, attempt, direct consequences, & reaction; 0 = not mentioned, 1 = vague representation of component, 2 = accurate representation, or 3 = verbatim representation of component*	▪ at post-training, experimental group recalled significantly more story components for both free recall and short-answer questions than control group ▪ at post-training, experimental group's note-taking summaries reflected significantly more main ideas than the control group ▪ at generalization, experimental group recalled significantly more story components for free recall and slightly more story components for short-answer questions than control group ▪ story grammar training produced significant improvements in recall and summarization of texts

			▪ generalization: free written retell & 10 short-answer questions		
Stein & Kirby (1992) Study was conducted in Canada	▪ 52 students ▪ conditions: (1) text absent or (2) text present	6th	▪ GMRT ▪ practice passage, practice written summaries, & discussion about summaries ▪ 1 expository passage ▪ 1-day & 1-week delay free oral retell	▪ participants read text ≥2 times ▪ wrote written summaries of text ▪ next day, asked tell as much as they could remember from original text ▪ recorded & transcribed ▪ 2 scorers ▪ *text analysis – propositions/ idea units included in retell; 0-points for M0 = propositions represent unimportant details, 1-point for M1 = propositions represent important details, 3-points for M2 = propositions represent main ideas, 5-points for M3 = propositions represent overall main idea or theme*	▪ interscorer reliability: 1-day recall = .95 & 1-week recall = .92 ▪ text present condition had a slightly greater recall (mean = 10.98) than text absent (mean = 9.54) ▪ high correlations between initial & delayed recall (*r*= .753) ▪ summary content was strongly related to recall in the text absent condition, but not in the text present condition ▪ summary depth was moderately related to recall in the text absent condition, but not in the text present condition ▪ reading ability was moderately related to recall in the text present condition, but not in the text absent condition ▪ reading ability was a significant predictor of recall ▪ deep summaries were produced in the text absent group only by more able readers
Zabrucky & Ratner (1992)	▪ 16 good comprehenders ▪ 16 poor comprehenders ▪ conditions: (1) context & target sentence congruent & close (2) context & target sentence congruent & far, (3) context & target sentence incongruent & close, or (4)	6th	▪ 8 narrative passages read on computer ▪ 8 expository passages read on computer ▪ free oral retell ▪ verbal report question: did the passage make sense? Why or why not? ▪ reading times for individual sentences	▪ participants read text on computer 1 sentence at a time, instructed to read for understanding, when felt understood sentence went on to the next sentence ▪ text was removed from view ▪ asked to tell all about the passage in their own words ▪ recorded & transcribed ▪ 2 scorers ▪ *text analysis – propositions/ idea units: a sequence of words or a word*	▪ interscorer agreement: retell protocols ≥90% ▪ good comprehenders recalled more context, target, & total idea units than poor comprehenders ▪ recall was greater for narrative than expository texts for both congruent & incongruent conditions ▪ good comprehenders recalled significantly more idea units in the incongruent condition than poor comprehenders ▪ good comprehenders were better able to verbally report on passage consistency

	context & target sentence incongruent & far		▪ # of look-backs or rereading for individual sentences ▪ oral decoding	*(modifier/connective) that conveyed a single idea; context, target, & total idea units recalled* ▪ *text analysis – proportion of correct idea units*	following reading ▪ students had significantly greater look backs for congruent expository passages than congruent narrative passages ▪ students had slightly greater look backs for incongruent expository passages than congruent narrative passages ▪ both good & poor comprehenders had significantly greater look backs for expository than narrative text ▪ students were more likely to reread expository sentences

Appendix B

Adapted RTF Record Form



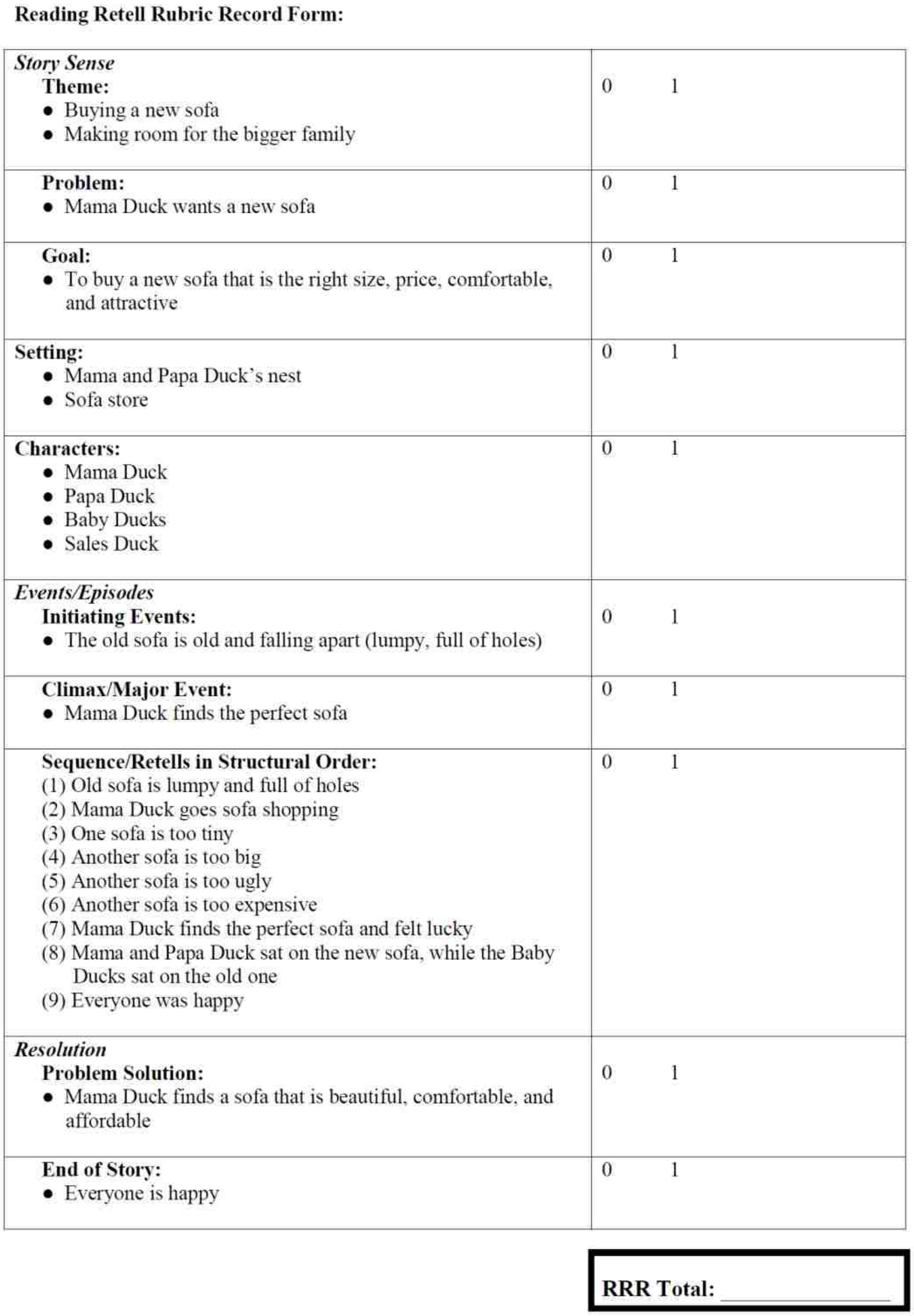
Retell Fluency Record Form:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120
121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140
141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160
161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180
181	182	183	184	185	186	187	188	189	180	191	192	193	194	195	196	197	198	199	200

RTF Total: _____________

Appendix C

Example of the Narrative RRR Record Form



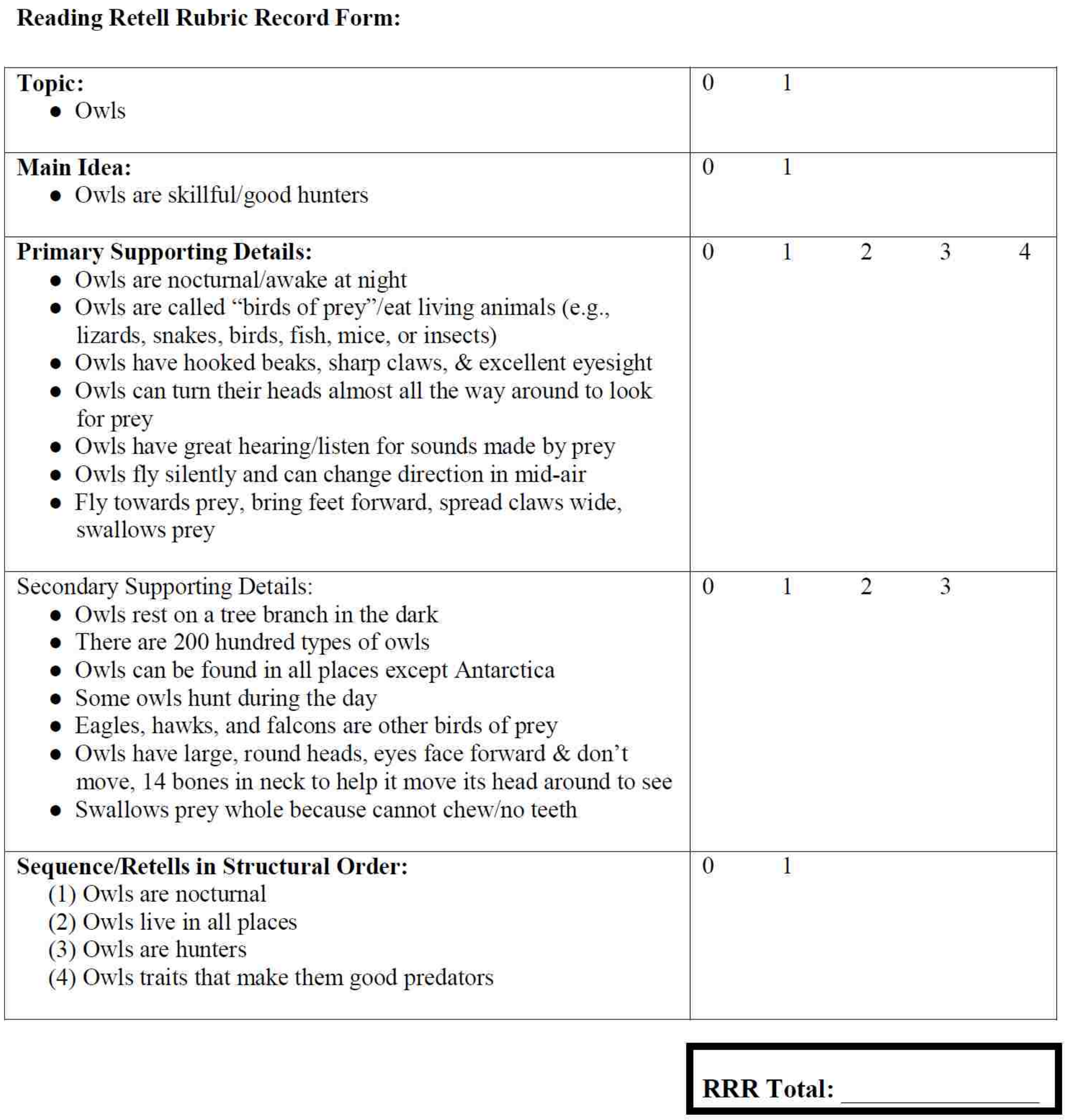
Reading Retell Rubric Record Form:

Story Sense **Theme:** ● Buying a new sofa ● Making room for the bigger family	0 1
Problem: ● Mama Duck wants a new sofa	0 1
Goal: ● To buy a new sofa that is the right size, price, comfortable, and attractive	0 1
Setting: ● Mama and Papa Duck's nest ● Sofa store	0 1
Characters: ● Mama Duck ● Papa Duck ● Baby Ducks ● Sales Duck	0 1
Events/Episodes **Initiating Events:** ● The old sofa is old and falling apart (lumpy, full of holes)	0 1
Climax/Major Event: ● Mama Duck finds the perfect sofa	0 1
Sequence/Retells in Structural Order: (1) Old sofa is lumpy and full of holes (2) Mama Duck goes sofa shopping (3) One sofa is too tiny (4) Another sofa is too big (5) Another sofa is too ugly (6) Another sofa is too expensive (7) Mama Duck finds the perfect sofa and felt lucky (8) Mama and Papa Duck sat on the new sofa, while the Baby Ducks sat on the old one (9) Everyone was happy	0 1
Resolution **Problem Solution:** ● Mama Duck finds a sofa that is beautiful, comfortable, and affordable	0 1
End of Story: ● Everyone is happy	0 1

RRR Total: _______________

Appendix D

Example of Expository RRR Record Form



Reading Retell Rubric Record Form:

Topic: ● Owls	0	1			
Main Idea: ● Owls are skillful/good hunters	0	1			
Primary Supporting Details: ● Owls are nocturnal/awake at night ● Owls are called "birds of prey"/eat living animals (e.g., lizards, snakes, birds, fish, mice, or insects) ● Owls have hooked beaks, sharp claws, & excellent eyesight ● Owls can turn their heads almost all the way around to look for prey ● Owls have great hearing/listen for sounds made by prey ● Owls fly silently and can change direction in mid-air ● Fly towards prey, bring feet forward, spread claws wide, swallows prey	0	1	2	3	4
Secondary Supporting Details: ● Owls rest on a tree branch in the dark ● There are 200 hundred types of owls ● Owls can be found in all places except Antarctica ● Some owls hunt during the day ● Eagles, hawks, and falcons are other birds of prey ● Owls have large, round heads, eyes face forward & don't move, 14 bones in neck to help it move its head around to see ● Swallows prey whole because cannot chew/no teeth	0	1	2	3	
Sequence/Retells in Structural Order: (1) Owls are nocturnal (2) Owls live in all places (3) Owls are hunters (4) Owls traits that make them good predators	0	1			

RRR Total: ________________

Lisa B. Thomas

EDUCATION

2004–Present

Doctoral Candidate in School Psychology
Subspecialization in Pediatric School Psychology
Lehigh University, Bethlehem, PA
APA accredited and NASP approved program

January 2007

M.Ed., Human Development
Lehigh University, Bethlehem, PA

December 2002

B.A., Psychology
B.S., Elementary and Kindergarten Education
The Pennsylvania State University, University Park, PA
The Phi Beta Kappa Society (2002)
Pi Lambda Theta – Honor Society in Education (2002)
Golden Key International Honour Society (2001)
Psi Chi – The National Honor Society in Psychology (2000)
The National Society of Collegiate Scholars (1999)

PROFESSIONAL CERTIFICATES

September 2010

Pennsylvania School Psychologist Certification
Lehigh University, Bethlehem, PA

December 2002

Pennsylvania Teacher Certification (K–6) Instructional I
The Pennsylvania State University, University Park, PA

June 1998

Reform Judiasm Teacher Certification
Jewish Community High School of Gratz College,
Melrose Park, PA

CLINICAL/PROFESSIONAL EXPERIENCES

2011–Present

Research Psychologist
Devereux Center for Effective Schools, King of Prussia, PA
Supervisor: Barry McCurdy, Ph.D., BCBA-D

2010–2011

Pre-doctoral Intern
Devereux Center for Effective Schools, King of Prussia, PA
University Supervisor: Christine Novak, Ph.D.
Site Supervisor: Amanda Lannie, Ph.D., BCBA-D

2006–2011 **Teacher/Psychometrician**
Sylvan Learning Center, PA
Supervisor: Vicky Henry, Director of Education

2007–2009 **School Psychology Practicum Student**
Lehigh Valley Hospital, Allentown, PA
University Supervisor: Patricia H. Manz, Ph.D.
Site Supervisor: Rosauro Dalope, MD

2007–2008 **School Psychology Practicum Student**
Whitehall-Coplay School District, Whitehall, PA
University Supervisor: Robin L. Hojnoski, Ph.D.
Site Supervisor: Kristin R. Stiles, Ph.D.

2006–2007 **Psychometrician**
Lehigh Psychological Services, Emmaus, PA
Supervisors: Adam J. Cox, Ph.D., ABPP and Daniel Werner, Psy.D

2006–2007 **School Psychology Practicum Student**
Allentown School District, Allentown, PA
University Supervisor: Robin L. Hojnoski, Ph.D.
Site Supervisors: Cheryl Bartholomew, Ed.S.

2005–2006 **Consultation Practicum Student**
Fountain Hill Elementary School, Bethlehem, PA
University Supervisor: Edward S. Shapiro, Ph.D.

2005 **Consultation Practicum Student**
A Charming World – A Division of Bright Horizons Family Solutions
Bensalem, PA
University Supervisor: Patricia H. Manz, Ph.D.

2002–2006 **In-Home Behavior Support Staff***, **Jamison, PA***
Access Services, Fort Washington, PA
Supervisor: Janet Roberto

2002–2005 **First Grade Hebrew School Teacher**
Old York Road Temple Beth Am Religious School, Abington, PA
Supervisor: Mimi Polin Ferraro, Director of Education

2003–2004 **Lead Teacher Two-year-olds, Pre-Kindergarten, & Kindergarten**
A Charming World – A Division of Bright Horizons Family Solutions
Bensalem, PA
Recipient of Bright Horizons' Spirit Award
Board member of the Better Together Committee
Supervisors: Michele Kelly and Teri Mayberry, Center Directors

2002–2003	**Substitute Teacher Kindergarten through Sixth Grade** ***Hatboro-Horsham School District, Lower Moreland Township School District, Norristown Area School District, and Upper Dublin School District, PA***
2002	**Second Grade Student Teacher** ***Handcock Elementary School, Norristown, PA*** *University Supervisor: Robin R. Headman, Ph.D.* *Site Supervisor: Amy Koerper, Classroom Teacher*
2001	**Third Grade Student Teacher** ***Frankstown Elementary School, Hollidaysburg, PA*** *University Supervisor: Linda K. Bergeman* *Site Supervisor: Pat Waring, Classroom Teacher*
2000–2001	**Psychology Intern** ***Progressions Northwestern Institute, Fort Washington, PA*** *University Supervisor: Keith A. Crnic, Ph.D.* *Site Supervisor: Barbara S. Hermey, Psy.D.*

Didactic Training Presentations

Thomas, L. (2012, February). *Introduction to curriculum-based assessment.* Pre-doctoral intern seminar training presented at the Devereux Foundation, King of Prussia, PA.

Thomas, L. (2012, February). *Group contingencies.* SW-PBIS universal team training presented at Webster Elementary School, Philadelphia, PA.

Blaze, T. & **Thomas, L.** (2011, October). *Group contingencies*. Staff training presented at Devereux Day School, Downingtown, PA.

Thomas, L., Ritvalsky, K., & Palmer, D. (2011, October). *Overview of PBIS.* School staff training presented at Devereux PA-CBHS Brandywine Campus, Glenmoore, PA.

Thomas, L. (2011, October). *SWIS and CICO.* Staff training presented at Thomas Ford Elementary School, Reading, PA.

Blaze, T., & **Thomas, L.** (2011, September). *Problem solving skills training*. Staff training presented at Devereux Day School, Downingtown, PA.

McCurdy, B., & **Thomas, L.** (2011, August). *School-Wide PBIS & Class-Wide PBIS.* Staff training presented at Devereux Day School, Downingtown, PA.

House, S. E., & **Thomas, L.** (2011, July). *Overview of positive classroom management strategies*. Staff training presented at Positive Outcomes Charter School, Dover, DE.

Lopez, J. C., & **Thomas, L.** (2001, June). *Overview of PBIS.* Clinical staff training presented at Devereux PA-CBHS Mapleton Campus, Malvern, PA.

Thomas, L., & Ritvalsky, K. (2011, May). *Overview of PBIS.* Residential staff training presented at Devereux PA-CBHS Mapleton Campus, Malvern, PA.

Thomas, L. (2001, May). *Increasing Active Engagement of Students through Providing Multiple OTRs.* Staff training presented at Devereux Day School, Downingtown, PA.

Lopez, J. C., & **Thomas, L.** (2011, March). *Overview of PBIS.* PBIS universal team training presented at Devereux PA-CBHS Brandywine Campus, Glenmoore, PA.

Thomas, L., & Ritvalsky, K. (2011, March). *Overview of PBIS.* PBIS school team training presented at Devereux PA-CBHS Mapleton Campus, Malvern, PA

Thomas, L. (2011, February). *PBIS team building.* PBIS universal team training presented at Devereux PA-CBHS Mapleton Campus, Malvern, PA.

Lopez, J. C., & **Thomas, L.** (2011, February). *Overview of PBIS.* PBIS universal team training presented at Devereux PA-CBHS Mapleton Campus, Malvern, PA.

Truckenmiller, A., **Thomas, L.**, & House, S. (2011, February). *SW-PBIS: New team training.* SW-PBIS universal team training presented at Reading High School, Reading, PA.

Truckenmiller, A., & **Thomas, L.** (2010, December). *Planning and implementing secondary SW-PBS: The Behavior Education Program (BEP)/Check-in Check-out (CICO).* SW-PBIS secondary team training presented at Webster Elementary School, Philadelphia, PA.

Truckenmiller, A., & **Thomas, L.** (2010, November). *PBIS Training.* SW-PBIS universal team training presented at Thomas Ford Elementary School, Reading, PA.

Lannie, A., & **Thomas, L.** (2010, November). *Thinking Functionally: Function-based intervention planning.* Staff training presented at Martin Luther School, Plymouth Meeting, PA.

Lannie, A., & **Thomas, L.** (2010, October). *Thinking functionally: Understanding student behavior.* Staff training presented at Martin Luther School, Plymouth Meeting, PA.

House, S., & **Thomas, L.** (2010, October). *AIMSweb: How to administer, score, and evaluate Reading-CBM.* Devereux Day School AIMSweb SWAT team training presented at Devereux Center for Effective Schools, King of Prussia, PA.

House, S., & **Thomas, L.** (2010, October). *Classroom management series: Classroom arrangement, routines and Premack principle.* Staff training presented at Devereux Day School, Downingtown, PA.

Thomas, L. B., McCurdy, E., & Seymour, K. (2010, March). *Developing effective positive behavior support plans.* Staff training presented at Bangor Area School District, Bangor, PA.

Thomas, L. B. (2009, March). *Behavioral Observation of Students in Schools (BOSS): Assessing the learning environment.* Staff training presented at Bangor Area School District, Bangor, PA.

Stiles, K., & **Thomas, L.** (2008, January). *Using AIMSweb to monitor students' response to intervention (RTI).* RTI Team training presented at Steckel Elementary School, Whitehall, PA.

Beck, M., Stiles, K., **Thomas, L**., & Seymour, K. (2008, January). *Dynamic Indicators of Basic Early Literacy Skills (DIBELS): How to administer, score, and evaluate.* DIBELS SWAT team training presented at Whitehall-Coplay School District, Whitehall, PA.

Thomas, L. (2005–2008). *Early Intervention for Young Children At-Risk for ADHD and Systematic Training for Effective Parenting (STEP; Dinkmeyer et al., 1997).* Parent education training program presented at Lehigh University, Bethlehem, PA.

RESEARCH EXPERIENCES

2010–2011 **Principal Investigator**
The effect of using response cards to increase opportunities to respond: The impact on student academic engagement and problem behavior
Devereux Foundation, Villanova, PA
Supervisor: Amanda Lannie, Ph.D.

2010–2011 **Research Assistant**
Development of Direct Academic Rating for Reading Comprehension
Center for Promoting Research to Practice
Lehigh University, Bethlehem, PA
Supervisors: Edward S. Shapiro, Ph.D.

2007–2010
Research Assistant
National Center on Response to Intervention
Center for Promoting Research to Practice
Lehigh University, Bethlehem, PA
Supervisors: Edward S. Shapiro, Ph.D.

2007–2010
Data Collection Coordinator
A Large Scale Study of Developing a Repeated Measure of Reading Comprehension
Center for Promoting Research to Practice
Lehigh University, Bethlehem, PA
Supervisors: Edward S. Shapiro, Ph.D. and Nanette S. Fritschmann, Ph.D.

2007–2008
Data Collection Coordinator
Examining Caregiver Variables in the Development of Early Number Sense in Young Children Project
Society of School Psychology sponsored grant
Lehigh University, Bethlehem, PA
Supervisor: Robin L. Hojnoski, Ph.D.

2005–2008
Principal Investigator
Predictors of social skills for preschool children at-risk for ADHD: The relationship between direct and indirect measurement
Qualifying Project (Master's thesis equivalent)
Lehigh University, Bethlehem, PA
Advisor: Edward S. Shapiro, Ph.D.

2004–2008
Data Collector/Consultant/Data Collection Coordinator/Database Manager
Early Intervention for Young Children At-Risk for ADHD
National Institute of Health sponsored grant
Lehigh University, Bethlehem, PA
Supervisors: George J. DuPaul, Ph.D. and Lee Kern, Ph.D.

2000–2002
Research Assistant
The Collaborative Family Study
National Institute of Child Health and Human Development sponsored grant
Child Study Center
The Pennsylvania State University, University Park, PA
Supervisors: Keith A. Crnic, Ph.D. and Craig Edelbrock, Ph.D.

2001
Research Assistant
Telephone Privacy Study
The Pennsylvania State University, Abington, PA
Supervisor: Peter B. Crabb, Ph.D.

2000–2001 **Principal Investigator**
The Pennsylvania State University, University Park, PA
Supervisor: Mary Napoli, Ph.D.
Title: How media and society impact our profession: Viewpoints from future classroom teachers

PUBLICATIONS

O'Dell, S. M., Vilardo, B. A., Kern, L., Kokina, A., Ash, A. N., Seymour, K. J., **et al.** (in press). JPBI 10 Years Later: Trends in Research Studies. *Journal of Positive Behavior Interventions.*

Thomas, L. B., Shapiro, E. S., DuPaul, G. J., Lutz, J. G., & Kern, L. (2011). Predictors of social skills for Preschool children at-risk for ADHD: The relationship between direct and indirect measurement. *Journal of Psychoeducational Assessment, 29,* 114–124.

Zirkel, P. A., & **Thomas, L. B.** (2010). State laws and guidelines for implementing RTI. *Teaching Exceptional Children, 43(1),* 60–73.

Zirkel, P. A., & **Thomas, L. B.** (2010). State laws for RTI: An updated snapshot. *Teaching Exceptional Children, 42*(3), 56–63.

Sokol, N. G., Kern, L., Arbolino, L. A., **Thomas, L. B.**, & DuPaul, G. J. (2009). A summary of home-based functional analysis data for young children with or at risk for Attention-Deficit/Hyperactivity Disorder. *Early Childhood Services: An Interdisciplinary Journal of Effectiveness, 3,* 127–142.

CONFERENCE PRESENTATIONS

Thomas, L. B. (2012, February). *Increasing active engagement of students with EBD through providing OTR.* Paper presented at the annual conference of the National Association of School Psychologist, Philadelphia, PA.

Thomas, L. B. (2012, February). *Evaluating a measure of reading comprehension for narrative and expository text.* Poster session presented at the annual conference of the National Association of School Psychologist, Philadelphia, PA.

Thomas, L. B. (2011, February). *Evaluating a measure of reading comprehension for narrative and expository text: Preliminary findings.* Poster session presented at the annual conference of the National Association of School Psychologist, San Francisco, CA.

Shapiro, E. S., Fritschmann, N. S., **Thomas, L. B.**, Hughes, C. L., & McDougal, J. (2010, August). *Preliminary development of a benchmarking measure for reading comprehension.* Poster session presented at the annual conference of the American Psychology Association, San Diego, CA.

Thomas, L. B. (2010, April). *Evaluating a Brief Measure of Reading Comprehension for Narrative and Expository Text: The Concurrent Criterion-Related and Predictive Validity of the Reading Retell Rubric.* Poster session presented at the Lehigh University College of Education Collaborative Mentoring Conference, Bethlehem, PA.

Thomas, L. B., Hojnoski, R. L., & Missall, K. N. (2009, February). *Caregivers beliefs and behaviors: Their relationship in early mathematics.* Poster session presented at the annual conference of the National Association of School Psychologists, Boston, MA.

DuPaul, G. J., Kern, L., **Thomas, L. B.**, Caskie, G., & Rutherford, L. E. (2009, February). *Early intervention for young children with ADHD: 24-month outcomes.* Paper presented at the annual conference of the National Association of School Psychologists, Boston, MA.

Thomas, L. B., Shapiro, E. S., DuPaul, G. J., & Lutz, J. G. (2008, February). *Predictors of social skills for Preschool children at-risk for ADHD: The relationship between direct and indirect measurement.* Poster session presented at the annual conference of the National Association of Psychologists, New Orleans, LA.

DuPaul, G. J., Kern, L., Arbolino, L. A., Booster, G. D., Hosterman, S. J., Kaduvettoor, A., **et al.** (2007, March). *Early intervention for preschoolers at-risk for ADHD: Strategies across settings.* Symposium presented at the annual conference of the National Association of School Psychologists, New York, NY.

Gilbert, K., Russell, J., & **Sigman, L.** (2001, November). *How media and society impacts our profession: Viewpoints from future classroom teachers.* Paper presented at the annual conference of National Council of Teachers of English, Baltimore, MD.

PROFESSIONAL MEMBERSHIPS

American Psychological Association, Division 16 (School Psychology)

Association of School Psychologists in Pennsylvania

National Association of School Psychologists
National Council of Teachers of English

Pennsylvania Psychology Association

PROFESSIONAL REFERENCES

Edward S. Shapiro, Ph.D.
Professor & Director
Center for Promoting Research to Practice
Lehigh University
Iacocca Hall, Room L-111-A
111 Research Drive
Bethlehem, PA 18015
Phone: 610.758.3258
Email: ed.shapiro@lehigh.edu

Barry L. McCurdy, Ph.D., NCSP, BCBA-D
Director
Devereux Foundation's Center for Effective Schools
2012 Renaissance Boulevard
King of Prussia, PA 19406
Phone: 610.542.3123
Email: bmccurdy@devereux.org