Lehigh University
Lehigh Preserve

Theses and Dissertations

2013

A Pilot Validation of an Academic Rating Scale of Reading Comprehension

Sarah Gebhardt
Lehigh University

Follow this and additional works at: http://preserve.lehigh.edu/etd

 Part of the Education Commons

Recommended Citation
Gebhardt, Sarah, "A Pilot Validation of an Academic Rating Scale of Reading Comprehension" (2013). *Theses and Dissertations.* Paper 1492.

This Dissertation is brought to you for free and open access by Lehigh Preserve. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Lehigh Preserve. For more information, please contact preserve@lehigh.edu.

A Pilot Validation of an Academic Rating Scale of Reading Comprehension

by

Sarah Gebhardt

Presented to the Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Doctor of Philosophy

in

School Psychology

Lehigh University

May 2013

Copyright by Sarah N. Gebhardt
2013

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 Director
Professor of School Psychology

_____
Accepted Date

Committee Members:

_____
Patricia H. Manz, Ph.D.
Associate Professor of School Psychology

_____
Mary Beth Calhoon, Ph.D.
Associate Professor of Special Education

_____
T. Chris Riley-Tillman, Ph.D.
Associate Professor of School Psychology,
University of Missouri

Acknowledgments

I would like to begin by expressing my appreciation and gratitude to my dissertation chair and academic advisor, Dr. Edward Shapiro, for the patient guidance and mentorship that he has offered during my time at Lehigh University. I am truly fortunate to have worked with an advisor who challenged me to continuously expand my knowledge and skills over my four years at Lehigh and allowed me to find my own path within the school psychology world. His personal commitment to all of his students' growth has been an inspiration.

I would also like to thank my dissertation committee, Dr. Patti Manz, Dr. Beth Calhoon, and Dr. Chris Riley-Tillman, for their time and their thoughtful feedback offered throughout the development of my dissertation project. Lehigh's program as a whole is an exceptionally welcoming environment that fosters a love of both the academic process and the field of school psychology, and my committee's mentorship has typified that experience. Similarly, I would like to extend my thanks to the faculty at my master's degree program at Miami University, particularly Drs. Steuart Watson, Kevin Jones, and Katherine Wickstrom; without their guidance I would never have seen potential in myself as a researcher or taken the leap and applied for a PhD program in the first place.

Last, but certainly not least, I would like to thank my loving family and friends who have been there to support me every step of the way. I would never have made it along this journey without the encouragement of my parents, Darren, and the many friends who have seen me through all of the emotional highs and lows of the last six years. Thank you for always being there to celebrate my achievements, however small, and to support me in times when I couldn't see the forest for the trees. I never would have gotten this far, personally or professionally, without you, and I am forever grateful.

TABLE OF CONTENTS

List of Tables

List of Figures

Abstract

Reading comprehension is a complex but critical skill that students must master to ensure future academic success. (e.g. Cain & Oakhill, 2007; Kintsch & Kintsch, 2005). Currently available reading comprehension scales are inadequate as classroom-based screeners for students struggling with reading comprehension; while some scales are psychometrically problematic, others are too lengthy and cumbersome to function well as a screener. (e.g. Keenan, Betjemann, & Olson, 2008). Research supports the validity of teacher-completed behavior rating scales of academic skills as an effective method of assessing students' academic progress that could address the current instrumentation gap in practice (e.g. Demaray & Elliot, 1998; Speece, Ritchey, Silverman, Schatschneider, Walker, & Andrusik, 2010). In response to noted weaknesses in the status of current reading comprehension screening assessments, a research team developed a rating scale of reading comprehension behaviors called the Rating Scale of Academic Skills – Reading Comprehension, Narrative (RSAS-RCN). This study conducted a pilot psychometric validation of the initial version of the RSAS-RCN measure. Participants included 41 general education teachers and 119 third, fourth, and fifth grade students across three school districts in eastern Pennsylvania. Students were administered the RSAS-RCN during the spring, and a small subset of students were re-administered the RSAS-RCN a week following the first administration. Students were also administered a standardized test of reading comprehension and a state standardized test of reading achievement. Results indicated that the RSAS-RCN is a highly reliable, single-construct scale and that the RSAS-RCN yielded consistent results over repeated administrations and a moderate to strong relationship with other tests of reading abilities. These results suggests that the RSAS-RCN may be a valuable new screening tool for reading comprehension measurement and are consistent with prior research

supporting the use of teacher-completed rating scales as a measure of students' academic skills. Implications for practice and future research are also discussed.

CHAPTER I. STATEMENT OF THE PROBLEM

In 2000, the National Reading Panel (NRP) published a report that listed five critical skills considered most vital in developing reading proficiency: phonemic awareness, phonics, fluency, vocabulary, and text comprehension. Due to the robust findings by the NRP (2000), these five skills have become the "gold standard" in reading instruction content. However, although much recent research has focused on the development of early intervention and assessment initiatives in basic literacy skills such as phonics and fluency, additional attention devoted to the development of the more complex, but equally vital, skill of reading comprehension is warranted (Cantrell, Almasi, Carter, Rintamaa, & Madden, 2010). It is estimated that at least 10% of children ages 7-11 years nationwide have specific reading comprehension deficits, unrelated to deficits in word-decoding skills (Cain & Oakhill, 2007; Nation & Snowling, 1997). If left unremediated, poor reading performance in elementary school is associated with continued academic failure throughout a student's academic experience (Juel, 1988). Effective identification, and thus effective assessment, of students struggling in the area of reading comprehension is therefore critical in order to intervene and assist these students in the school setting (NRP, 2000).

Reading Comprehension

Cognitive Skills. Reading comprehension is often understood as a complex process of synthesizing cognitive skills, which include visual perception of text, ability to hold text in working (short-term) memory during decoding, construction of mental schemas ("pictures" of what text is describing), and retrieval of prior knowledge from long-term memory storage, in order to engage in reading and extract meaning from text (Kintsch & Kintsch, 2005). Kintsch (2004) describes a construction-integration model of reading comprehension, which

conceptualizes comprehension as a network of processes occurring at and across three different levels of cognition that work together to produce mental representations of text. The first level consists of decoding processes, wherein students utilize print recognition and phonological processing skills to identify words and organize them into brief "propositions," or idea units. Processes involved at this stage include visual perception of the text, holding the text in working memory during recognition and decoding of individual words, and joining words into phrases to begin extracting meaning. At the secondary level, propositions are synthesized into two levels of idea networks, known as the text microstructure and the text macrostructure. At the microstructure level, syntactic (sentence-level) information is processed to form complete idea propositions that represent the full meaning of the text. At the macrostructure level, the organization of these ideas, for example into setting, climax, and denouement, is recognized.

Finally, situational models, or mental models of the text, are constructed through the integration of the microstructure and macrostructure of the text with background information from the reader's prior knowledge. Situational models are influenced by other cognitive abilities, including working and long-term memory capacity and inference making ability. Working memory and short-term memory are required to hold the representation of the current text, while long-term memory activation is required in order to retrieve prior knowledge that will further inform the situational model. Inference-making requires readers to again utilize working and long-term memory to extrapolate what they have gleaned from a text so far and create a new mental model connecting ideas that may not be explicitly stated in the text (Kintsch, 2004).

Metacognitive Skills. Reading comprehension is additionally supported by students' use of metacognitive strategies to monitor and control their own progress towards their goal of understanding a text (Hacker, 2004). Hacker (2004) proposes a model of self-regulated reading,

that he conceptualizes as an interaction between cognitive and metacognitive processes, which include goal setting, models of prior experience, metacognitive experiences during comprehension, and monitoring and control strategies. Monitoring strategies include re-reading, looking back to prior text, and predicting, while control strategies can include summarizing and referencing other texts to clarify or correct understanding of current text. At each level of cognitive processing, such as the three levels proposed in Kintsch's (2004) model, readers can evaluate and compare the internal representation that they are constructing (cognitive model) of the current text with pre-existing (metacognitive) models of prior experiences at that same level, whether models are at the decoding, microstructure, or situational model level. Readers can use monitoring strategies (e.g. re-reading) to check their understanding of text as they read, and if they note differences between their understanding of current text and their understanding of prior knowledge, they can use control strategies (e.g. referencing prior text) to correct their current understanding. Metacognitive strategies therefore help reduce demands on memory and increase opportunities for information processing throughout comprehension (Hacker, 2004).

Research has further confirmed the integral nature of metacognitive processes during comprehension (Cain, Oakhill, & Bryant, 2004). As part of a longitudinal study examining the relationship between working memory and reading comprehension skills, one study examined the contributions of inference making, comprehension monitoring, and story structure knowledge to reading comprehension (Cain et al., 2004). A sample of 102 typically-developing English children were assessed at 8, 9, and 11 years of age on standardized measures of reading comprehension, word reading accuracy, vocabulary, verbal ability, and working memory, and on researcher-developed measures of inference making, comprehension monitoring, and story structure knowledge. Results of a hierarchical linear model controlling for word reading

accuracy, vocabulary, verbal ability, and working memory found that comprehension monitoring and inference making ability still contributed unique and significant variance to reading comprehension, beyond the control variables (Cain et al., 2004).

Differentiation Between Successful and Unsuccessful Comprehenders. Research supports that both higher-order cognitive and metacognitve skills can also effectively differentiate between students who experience problems related to reading comprehension from those who do not. For instance, research has indicated that children who have difficulty with comprehension, or "poor comprehenders," may struggle specifically with the higher order cognitive comprehension skills in comparison to their "successful comprehender" peers. For example, in a study examining 187 children ages 7-10 years, Nation and Snowling (1997) found that while children identified as poor comprehenders performed significantly more poorly than their successful comprehender peers on tests of linguistic comprehension including single word reading, text level reading, and comprehension, the two groups did not significantly differ on an assessment of word decoding ability. Research further suggests that use of metacognitive skills may aid in distinguishing which students are poor comprehenders from successful comprehenders (Palincsar & Brown, 1987; Garner & Kraus, 1981-1982). For instance, in one review of predictors of poor comprehenders, both comprehension monitoring and knowledge of story structure, a metacognitive model, were identified as effective predictors; research has indicated that poor comprehenders are less likely to notice text inconsistencies and have more difficulty identifying different story elements than their more successfully comprehending peers (Cain & Oakhill, 2007).

Skill Development. Growth models of reading comprehension skills help further elucidate the picture of students' comprehension at different stages of development, and have

indicated that the development of skills needed for comprehension progresses in complexity over the course of development (Johnston, Barnes, & Desrochers, 2008). For instance, a longitudinal study of a racially diverse sample of 626 children, all from families who qualified for participation in Head Start, examined the development of their reading abilities from preschool through the 4th grade. Results indicated that in early childhood, reading comprehension is best predicted by early literacy and word decoding skills, such as letter knowledge and phonological awareness, but by 3rd and 4th grade, reading comprehension was best predicted by more complex oral language skills including vocabulary processes, syntactic ability, and narrative recall (Storch & Whitehurst, 2002). Thus, using Kintschs' (2004) model, young children conduct most of their comprehension at the decoding level, while older children can begin to process microstructure, macrostructure, and situational models as they develop.

Current Measures of Reading Comprehension

Reflective of the highly complex nature of reading comprehension, methods of reading comprehension assessment have evolved over years in both format and content (Pearson & Hamm, 2005). Some reading comprehension tests, such as cloze passages wherein a student fills in deleted words in a passage, have focused on decoding and micro level skills, while others, such as question answering items that can include both multiple choice and open-ended questions, can address macrostructure and situational models (Pearson & Hamm, 2005). Despite this broad array of comprehension assessment methodologies that have been present over the years, the current status of reading comprehension instrumentation continues to be variable and the need for improvement is clear (Sweet, 2005).

Evidence suggests that measures of reading comprehension vary widely in the contribution of word decoding vs. oral language comprehension skills to their outcome scores,

and results therefore may not correlate well across measures (Keenan, Betjeman, & Olson, 2008). For instance, Cutting and Scarborough (2006) conducted a study examining the relative contributions of reading, language, and cognitive skills to three widely used, standardized measures of reading comprehension (the Gates-MacGinitie Reading Test (G-M; MacGinitie, MacGinitie, Maria, & Dreyer, 2000), the Gray Oral Reading Test – Third Edition (GORT-3; Wiederholt & Bryant, 1992), and the Wechsler Individual Achievement Test – Passage Comprehension subtest (WIAT; Wechsler, 1992)). Using a sample of 97 children, from grades 1 – 10, the study examined relative contributions of word decoding skills and oral language comprehension constructs skills to each reading comprehension measure, the contributions of other factors including reading rate, verbal memory, and attention, and also correlations across the three measures.

Results of hierarchical regression analysis indicated that overall word decoding skills and oral language comprehension skills accounted for significant variance across all three reading comprehension measures, but relative contributions varied across the three measures. For instance, while lexical (single word) and sentence processing both made significant unique contributions to G-M variance, only lexical processing significantly contributed to GORT-3 variance, while only sentence-level processing made a significant contribution to WIAT variance. Additionally, 1-6% of variance was accounted for by reading speed in all three comprehension measures, but no other variables such as verbal memory, were significant predictors of scores on the G-M, GORT-3, or WIAT. Correlations of total scores across the measures also varied; correlation between the G-M and WIAT was in the high range (r = .79), but correlations were considerably lower between the GORT-3 and the WIAT (r = .70) and between the GORT-3 and the G-M (r = .64). These results highlight the lack of content

consistency that characterizes current reading comprehension measurement, and also emphasizes the need for improved methods of reading comprehension assessment.

Given the uncertain validity of standardized tests of reading comprehension and the sometimes cumbersome nature of administering many such tests, researchers have also examined other, more brief measures of reading skill that might be used as indicators of reading comprehension skill (Hosp & Fuchs, 2005). To illustrate, Fuchs, Fuchs, and Maxwell (1988) investigated the validity of four brief reading measures (oral reading fluency (ORF), passage retell, question answering, and cloze) compared with the Stanford Achievement Test – 7th edition Word Study Skills subtest (SAT-WS) and Reading Comprehension subtest (SAT-RC; Gardner, Rudman, Karlsen, & Merwin, 1982). Results indicated that the ORF measure was most highly correlated to SAT-RC scores (r = .91), while the ORF, passage retell, and question answering measures all demonstrated construct validity, as defined by the researchers (significantly more correlated to the SAT-RC than the SAT-WC). Other studies have provided further evidence of the strong relationship between measures of ORF and standardized measures of reading comprehension (Skinner, Williams, Morrow, Hale, Neddenriep, & Hawkins, 2009; Marcotte & Hintze, 2009). Some researchers, however, question the validity of ORF and other such brief reading measures as indicators, both because such scales are not actually measuring comprehension and may not be causally linked, and because analyses comparing ORF and reading comprehension often do not statistically account for the floor and ceiling effects associated with ORF, which may lead to misinterpretations of correlational data (Paris, Carpenter, Paris, & Hamilton, 2005). Thus, the reviewed research suggests that neither current standardized, summative tests of reading comprehension nor brief, formative evaluations of

reading skills lead to consistent, complete and statistically validated pictures of students' reading comprehension skills.

As evident from the discussion above, the development of effective reading comprehension assessment is an issue that remains largely unresolved in the literature. Assessment design, content, and utility vary widely and reflect varying components of reading comprehension, and some lack sound psychometric qualities. Furthermore, despite research supporting the critical nature of both cognitive and metacognitive skills in reading comprehension, there is not yet an assessment of reading comprehension that incorporates assessment of these observable metacognitive skills.

Behavior Rating Scales

An appropriate method for such an assessment that is not currently being utilized to measure reading comprehension might be the use of behavior rating scales. Behavior rating scales are typically standardized instruments that ask a rater (e.g. a teacher) to answer a series of items rating a ratee's (e.g., a student's) behavior along a continuum (e.g. 1 = never, 5 = always). Behavior scales are considered indirect measures because they ask a rater to make retrospective, summative judgments about a ratee's behavior over a long period of time, but they are also considered objective because, when psychometrically valid, rating scales reliably measure constructs over time, compare outcomes to standardized norms, and are meaningfully related to other measures of behavior constructs (Merrell, 2000).

Teacher Judgment. One of the primary advantages that the use of behavior rating scales in measuring reading comprehension could offer is that behavior rating scales would actively utilize teacher judgment of reading skills. Historically, research has supported the validity of teacher-based measurements of academic skills, particularly in the area of reading achievement

(e.g. Demaray & Elliot, 1998). In a landmark research review, Hoge and Coladarci (1989) examined 16 studies related to teacher judgments of academic achievement. Studies reviewed included those that correlated teacher judgments of student achievement levels using likert-type scales with students' actual standardized test scores, characterized as "indirect" studies, and studies that correlated teachers' specific predictions of standardized test scores with students' actual standardized test scores, characterized as "direct" studies. Results indicated that teacher judgments fell at the high end of the moderate range overall, with the median correlation for direct studies (r=.69) slightly higher than the median correlation for indirect studies (r=.62).

Research has continued to explore the relationship between teachers' judgments of student achievement and student performance on standardized measure of achievement. For instance, Demaray and Elliot (1998) examined the accuracy of teacher ratings of overall academic competence using the Academic Competence subscale of the Social Skills Rating System –Teacher (SSRS-T; Gresham & Elliott, 1990) to total scores on the Kaufman Test of Educational Achievement (K-TEA; Kaufman & Kaufman, 1985). The Academic Competence scale of the SSRS-T includes four questions pertaining to reading and five pertaining to overall academic performance, motivation, parental encouragement, intellectual functioning and classroom behavior. Participants included 12 first through fourth grade teachers and 47 first through 4th grade students. Teachers rated students using the Academic Competence scale and also put a '+' by items on the KTEA Math, Reading, and Spelling subtests they believed students would answer correctly and a '-' next to items they believed students would answer incorrectly. Students were administered the K-TEA Math, Reading, and Spelling subtests which were used to calculate a total achievement score. Results indicated that overall correlations between teacher ratings on the SSRS-T Academic competence scale and the students' actual K-TEA total score

were high (r=.70), and overall correlations between teachers' predicted K-TEA scores were also high (r=.84) and the median percentage of agreement between student responses on individual K-TEA items and teachers' predicted responses to K-TEA items was 79%.

Other studies have also examined teacher accuracy in judging student reading performance on curriculum-based measures. For instance, one study asked two teachers to predict 33 first, second and third grade students' instructional reading level and class-wide comparison of students' reading ability across students using a ranking system (Eckert, Dunn, Codding, Begeny, & Kleinman, 2006). Teacher predictions of instructional levels (e.g., Mastery, Instructional, Frustrational) were then correlated with students' actual instructional level as indicated by their oral reading fluency data. Results indicated moderate to high correlation between teacher judgment of instructional level and actual instructional level (first grade r = .59, second grade r = .72, third grade r = .83). Another study by Feinberg and Shapiro (2003) examined the ability of 30 third, fourth, and fifth grade teachers to predict reading achievement using the Academic Competence Evaluation Scales (ACES; DiPerna & Elliot, 1999) and estimate students' oral reading fluency. Teacher predictions were then correlated with actual oral reading fluency scores. Results indicated moderate correlation between predicted and actual oral reading fluency rates (r=.70) and between ACES and actual oral reading fluency rates (r= .62) and found that, overall teachers were more accurate in recognizing relative differences in strength across students in a classroom than at estimating specific levels of oral reading fluency.

Measuring Academic Skills. Behavior rating scales also have the advantage of a history of successful use in the school setting for measuring both internalizing and externalizing behavior constructs, including social skills, aggression, hyperactivity, and anxiety (Gresham, 2007). However, although rating scales for behaviors such as aggression and hyperactivity are

widely used in the school setting, rating scales for skills directly related to academic performance are less common. There is, however, some basis for teacher rating scales of academic skills in the literature, as illustrated, for instance, by some brief, researcher-generated scales present in the teacher judgment literature (e.g. Begeny, Krouse, Brown, & Mann, 2011).

There are also several examples of ratings scales of overall academic competence that exist in the literature and have been utilized in the school setting. One such example is described in a validation study of the Academic Competence Evaluation Scales (ACES; DiPerna & Elliot, 1999). The ACES is a 95-item rating scale that consists of five proposed components on which teachers rate individual students: academic skills, study skills, academic motivation, interpersonal skills, and academic self-concept. Researchers asked 56 elementary teachers (Grades 1^{st}-6^{th}) to rate 300 students on the ACES and, for a subset of the students, on the two criterion measures, the Social Skills Rating System (SSRS; Gresham & Elliott, 1990), a behavior rating scale that consists of three components including social skills, problem behaviors, and academic competence, and the Iowa Test of Basic Skills (ITBS; Hoover, Hieronymus, Frisbie, & Dunbar, 1993), a standardized test of achievement. Results indicated that the five proposed components for the ACES were supported by exploratory factor analysis and that all five components demonstrated high internal consistency (.92-.98) and adequate test-retest reliability (.70-.92). Further, correlations with the ITBS total score were moderate to high (.52-.84), as were correlations with the SSRS academic competence scale (.43-.87), suggesting that teachers were able to estimate individual students' academic abilities relative to measures of achievement with moderate to high accuracy, and that these estimates were consistent across rating scales.

Similarly, DuPaul, Rapport, and Perriello (1991) also developed and validated a 19-item behavioral rating scale intended to identify the presence of academic skills deficits in the

classroom, specifically for students who also have disruptive behavior disorders, the Academic Performance Rating Scale (APRS). DuPaul, Rapport, and Perriello assessed a sample of 493 students (grades 1st-6th) for factor analysis and a subsample of 60 students for reliability and validity confirmation analyses using the APRS, two brief measures of Attention Deficit/Hyperactivity Disorder (ADHD) symptoms, an academic efficiency scores (AES), which the researchers calculated by determining the number of items on a school assignment accurately completed by the target child during a set observational period, and on the Comprehensive Test of Basic Skills (CTB; CTB/McGraw-Hill, 1982), a school-based, norm-referenced achievement test that includes subscale on math, reading, and language. Results supported the presence of three scale components, Academic Success, Impulse Control, and Academic Productivity, and indicated adequate internal consistency (.72-.95) and high test-retest reliability (.88-.95) across the components and scale total score. Validity coefficients indicated moderate correlations between the APRS total and all criterion measures (.48-.72), with slightly higher correlations between the CTBS subscales and the APRS academic success component (.61-.62) than with the APRS total score (.48-.53). These results further support that teacher observations may be useful to estimate academic performance in the classroom, even amongst students who exhibit significant disruptive behaviors.

In addition, a very small body of research has begun to investigate the possibility of incorporating teacher rating scales of academic skills into screening batteries for learning problems (Speece, Ritchey, Silverman, Schatschneider, Walker, & Andrusik, 2010). For example, one study examined the ability of kindergarten teachers to identify students with early learning problems using a researcher-generated rating scale (Taylor, Anselmo, Foreman, Schatschneider, & Angelopoulos, 2000). A total of 303 kindergarten students were assessed for

potential learning problems using a comprehensive battery that included overall cognitive ability assessment by the Kaufman Assessment Batter for Children (K-ABC; Kaufman & Applegate, 1988) and academic achievement assessed by the Peabody Individual Achievement Test-Revised (PIAT-R; Markwardt, 1989) Reading Recognition, Spelling, and Mathematics subtests. Parents and teachers also completed multiple behavior rating scales including the Child Behavior Checklist (CBCL; Achenbach, 1991) and the Conners Hyperactivity Checklist (Barkley, 1990). Finally, teachers also completed a researcher-generated teacher rating scale of student performance on six critical kindergarten academic skills (letter naming, letter sounds, correspondence of letters and words with oral language, naming numerals 1 to 10, counting with 1-to-1 correspondence from 1 to 10, and matching numerals with object sets from 1 to 10). Results of the study indicated that students identified as at-risk by the teacher rating scale scored significantly lower on overall cognitive and achievement measures and on teacher-completed ratings of behavior than students not identified as at risk. Also, among a sub-sample of identified and not identified kindergarten students re-tested with the PIAT-R during 1st grade, identified students continued to score significantly lower across achievement areas than not identified students. The researchers argue that these findings support the validity of teacher judgments of early academic skills and demonstrate the potential utility of teacher ratings as a screener for learning problems.

Although behavior rating scales have a clear history of successful use in schools for assessing behavior constructs such as hyperactivity and academic competence, and a few brief rating scales of general academic competence and researcher-generated brief scales of reading behavior exist, no fully validated, in-depth rating scales of academic constructs such as reading

comprehension have yet entered the literature. Thus, the potential utility of teacher-completed rating scales of academic behaviors in the classroom settings remains unexplored.

Direct Behavior Rating Scales

Another observational rating method, known as direct behavior rating (DBR), may provide an advantageous format to supplement traditional rating scales, particularly for the purpose of progress monitoring. As detailed by Chafouleas and colleagues (2009), DBR combines desirable characteristics of both traditional rating scales and systematic direct observation (SDO), a highly accurate but time- and training-intensive methodology wherein a rater observes a ratee and records instances of a target behavior as it occurs in real time.

Chafouleas and colleagues describe DBR's three defining characteristics as being that DBR is direct, because it is conducted at the time and place a behavior occurs, DBR targets observable behavior, and DBR involves rating, or quantifying a rater's perception of a target individual's behavior. Current evidence is emerging for validation DBR scales across a variety of behaviors and age groups, including preschool, kindergarten, and middle school (Chafouleas, Christ, Riley-Tillman, Briesch, & Chanese, 2007; Briesch, Chafouleas, & Riley-Tillman, 2010; Chafouleas, Briesch, Riley-Tillman, Christ, Black, & Kilgus, 2010). All DBR studies to date have used DBR to assess behaviors such as academic engagement and disruptive behavior, but no literature to date has examined the use of DBR methodology for assessing academic skills.

The Rating Scale of Academic Skills – Reading Comprehension, Narrative

In sum, reading comprehension is a complex but critical skill that students must master to ensure future academic success. Currently, the available reading comprehension scales are psychometrically problematic and tend to be either too lengthy or cumbersome in administration to function well as a screener. Additionally, in the case of some formative assessments, some

scales are too brief to provide a comprehensive picture of a student's reading comprehension skills. A teacher-completed behavior rating scale of academic skills, including reading comprehension may represent an effective method of assessing students' academic progress, and would offer the opportunity to utilize teacher judgment. However, at this time, no validated rating scale of reading comprehension behaviors exists.

In response to these gaps in both the reading comprehension assessment literature and the lack of existing rating scales for academic behaviors utilizing teacher judgment in the literature, Dr. Edward S. Shapiro and colleagues at Lehigh University's Center for Promoting Research to Practice recently developed an academic rating scale measure of reading comprehension, the Rating Scale for Academic Skills – Reading Comprehension, Narrative (RSAS-RCN). The RSAS-RCN was conceptualized as a broad-based, brief, efficient screening tool for teacher use in the classroom setting to help identify students struggling with reading comprehension skills. The RSAS-RCN items were designed to examine highly specific reading comprehension skills to aid in linking assessment outcomes to instruction. If the RSAS-RCN is found to be a valid measure of reading comprehension, it is hoped that individual items will be able to be taken out and used as direct, single-skill rating scales in a manner similar to DBR procedures to monitor student progress in specific areas of reading comprehension. Before that step can be addressed, however, the RSAS-RCN measure as a whole must undergo rigorous psychometric validation.

The purpose of this study is to conduct a pilot psychometric validation of the initial iteration of the RSAS-RCN measure. This psychometric validation will be approached using both a historical view of test validation, which emphasizes examination of the relationship among test items and with other tests of similar constructs, and by using Messick's (1995) more contemporary view of construct validity. Messick's (1995) theory posits that there are six

primary dimensions of construct validity: 1) the content aspect, which includes evidence of the relevance and representativeness of the scale items as appraised by expert professional judgment; 2) the substantive aspect, which refers to empirical evidence of the representativeness (including internal reliability estimates) of the scale items to the measures' construct; 3) the structural aspect, which refers to the fidelity or rational consistency of the scale's scoring method to the construct it measures; 4) the generalizability aspect, which includes how well the measure's score generalizes across raters, rates, settings, and administrations; 5) the external aspect, which refers to empirical evidence of the measure's relationship to other measures of the same construct; and 6) the consequential aspect, which includes consideration of the actual and possible social consequences of administering the measure, with an emphasis on potential sources of bias or unfairness in scoring and interpreting the measure. With these aspects of validity in mind, this study examined the psychometric properties of the RSAS-RCN by addressing the following research questions:

1) Is there adequate evidence for the construct validity of the RSAS-RCN, including substantive validity and structural validity?

2) What is the test-retest reliability of the RSAS-RCN total score?

3) What is the external validity of the RSAS-RCN with a standardized assessment of reading comprehension (e.g. GRADE)?

4) What is the diagnostic validity (i.e. sensitivity, specificity, negative predictive power, and positive predictive power) of the RSAS-RCN to levels of reading proficiency as determined by a state-wide, standards-based assessment of reading (e.g. PSSA-Reading)?

5) What is the social validity of the RSAS-RCN to teachers, as measured by an informal, teacher-completed questionnaire pertaining to acceptability?

CHAPTER II. LITERATURE REVIEW

Reading comprehension is a critical but complex skill that students must master to ensure continued academic success (NRP, 2000). Reading comprehension is often conceptualized as a process of constructing mental models from text through the use of both cognitive skills, such as working memory, information processing, and long-term memory retrieval, and metacognitive skills, such as monitoring and control strategies for self-regulation of comprehension (Hacker, 2004; Kintsch, 2004). Given the complexity of the reading comprehension process, it is not surprising that at least 10% of students nationwide struggle with specific learning deficits in reading comprehension (Cain & Oakhill, 2007). In order to identify and provide intervention to these struggling students, effective reading comprehension assessments must be available to practitioners. Currently, however, the literature on reading comprehension assessment indicates need for improvement.

Current Reading Comprehension Measures

Reading comprehension assessment has evolved throughout the 20th century both in the types of skills that are assessed and format of tests used (Pearson & Hamm, 2005). Yet in spite of years of research and assessment iterations, reading comprehension assessment remains a subject of debate in the academic field. A review of current comprehension assessment literature reveals that there are many unresolved issues regarding effective assessment of students in schools, including both psychometric problems and issues of utility in the classroom (Sweet, 2005).

Skills Assessed. A primary concern regarding current standardized and norm-referenced tests of reading comprehension is that such tests vary widely in the skills they assess and the weight different skills are given in describing a student's performance on a test. For instance, Nation and Snowling (1997) examined the relative contributions of single word-reading and

listening comprehension to two British tests of reading comprehension, the Neale Analysis of Reading Ability-Revised (NARA-II; Neale, 1989) and the Suffolk Reading Scale (Suffolk; Hagley, 1987). The researchers first conducted an exploratory factor analysis (EFA) of word reading and listening comprehension factors on the NARA-II and Suffolk and determined that the two comprehension scales loaded on both the word decoding and listening comprehension factors, but the Suffolk loaded much higher on word decoding while the NARA-II loaded primarily on listening comprehension. The researchers then followed up with a Hierarchical regression of listening comprehension and word decoding to the two scales and found that when listening comprehension was entered as Step 1, listening comprehension predicted 44% of variance on the NARA-II and only 17% of variance on the Suffolk, while word reading predicted only 25% of variance on the NARA-II but 62% of the variance on the Suffolk (Nation & Snowling, 1997).

Similarly, Keenan, Betjemann, and Olson (2008) examined the relative contributions of word decoding and listening comprehension to five different measures of reading comprehension, the Gray Oral Reading Test-3 (GORT-3; Wiederholt & Bryant, 1992), the Qualitative Reading Inventory-3 Questions and Retell subtests (QRI-Questions and QRI-Retell; Leslie & Caldwell), the Woodcock-Johnson Passage Comprehension subtest (WJPC) from the Woodcock-Johnson Tests of Achievement-III (WIAT; Woodcock, McGrew, & Mather, 2001), and the Reading Comprehension subtest from the Peabody Individual Achievement Test (PIAT; Dunn & Markwardt, 1970). Participants for this study were 510 children between 8 and 18 years old. As in Nation and Snowling (1997), all reading comprehension tests loaded on both word decoding and listening comprehension factors in an exploratory factor analysis, but the WIAT and the PIAT loaded much higher on word decoding than listening comprehension, and these

findings were again confirmed by hierarchical regression modeling. In addition, the researchers conducted correlation analyses across the five comprehension tests and found that all inter-test correlation range were in the low to moderate range (.31-.54) except for one high correlation between the WIAT and the PIAT (.70), which the researchers hypothesize is attributable to the high degree of word-decoding content in those tests (Keenan et al., 2008).

The disparity in skill contributions described by these studies indicates that reading comprehension tests are not uniform in the constructs they assess. The major implication is that, depending on individual patterns of strengths and weaknesses across word decoding and listening comprehension skills, an individual student may not score similarly on different standardized comprehension assessments, even though the tests all supposedly measure the same "reading comprehension" construct. A further implication of these results is that skill instruction resulting from outcomes on these tests vary depending on the test used. Instructors and researchers have also examined possible reasons within the tests themselves that might indicate why disparities in skill contributions persist across assessments.

Item Formats. One variable that can potentially impact outcomes are the format of items on tests of comprehension. Spear-Swearling (2004) illustrated the differences in skills assessed across item formats in her investigation of fourth-graders' on two different state-mandated tests of reading comprehension. Both tests were part of the Connecticut Mastery Test (Connecticut Department of Education, 2000). The first test used a cloze item format, wherein students fill in missing words in sentences, and the other used a question-answering format, wherein students answered both open-ended and multiple choice questions about passages they read. Using a socio-economically diverse sample of 95 fourth grade students, hierarchical linear regression analysis indicated that, when word accuracy was entered as the first step of analysis,

word accuracy (e.g. decoding-level skill) accounted for 51% of the variance in the cloze test and 34% of the variance in the question-answering test, while oral comprehension accounted for 14% of the variance in the cloze test and 21% of the variance in the question-answering test, suggesting that the cloze test was much more influenced by word-level reading skills than the question-answering test (Spear-Swerling, 2004). Similarly, Nation and Snowling (1997) hypothesized that one factor accounting for the much larger contribution of word-level reading skills than listening comprehension skills to the Suffolk than the NARA-II is that the Suffolk uses a cloze format, while the NARA-II uses a question answering format.

Item Difficulty. Another important consideration is the level of difficulty of test items and the level of inference such items require. For instance, Keenan and colleagues (2008) suggest that their finding that both the PIAT, which uses a multiple-choice question-answering format, and the WJPC Passage Comprehension subtest, which uses a cloze format, have the majority of their variance accounted for by word decoding skills is that both tests require students to read very short (1-2 sentence) passages, while the other tests included in their study, such as the Qualitative Reading Inventory (QRI; Leslie & Caldwell, 2001), used substantially longer passages. Two other studies have specifically examined the effects of passage difficulty on the GORT-3 and GORT-4 (Weiderholt & Bryant, 2001). Keenan and Betjemann (2006) used a sample of 77 undergraduates and a secondary sample of 10 children ages 7-15 and administered the GORT-3 questions without allowing participants to see the passage the questions were about. Results indicated that undergraduate respondents were able to answer 86% of GORT-3 passageless questions with above chance accuracy, while the children were able to answer 47% of questions with above chance accuracy. A second study investigated the utility of the GORT-4 with a group of adults with low (3^{rd}-5^{th} grade level) literacy proficiency (Greenberg,

Pae, Morris, Calhoon, & Nanda, 2009). Results indicated a serious problem with passage scaling; 50% of participants reached a ceiling after answering questions for story 1 or story 2, but if standard administration was ignored and administration continued, these same participants were able to establish a basal on stories 3 and 4. This suggests that the first and second stories are more difficult than those following, and if standard administration is followed, the GORT-4 may underestimate the actual skill level of many readers (Greenberg et al., 2009). Finally, in another study, Bowyer-Crane and Snowling (2005) examined the level of inference required by two question-answer format reading comprehension tests, the NARA-II and the Wechsler Objective Reading Dimensions Test of Reading Comprehension (WORD; Wechsler, 1990). A qualitative analysis of the content of items from both tests by several literacy experts determined that the WORD items required more literal and elaborative (e.g. lower level) inferences and the NARA-II required more cohesive and knowledge-based (e.g. higher level) inferences.

Group Differentiation. Clearly, a serious consequence of the variations in skills and formats of standardized reading comprehension tests is that these tests, then, also vary in the way they differentiate between struggling and successful comprehenders. For example, in a second phase of Bowyer and Snowling's (2005) study, a group of 10, 2^{nd}-6^{th} grade skilled comprehenders and a group of 10, 2^{nd}-6^{th} grade less skilled comprehenders were assessed using the WORD and the NARA-II. Results indicated that although the skilled group performed significantly better than the unskilled group on both measures, there was also a group by scale interaction which indicated that the less skilled group performed significantly better on the WORD than the NARA-II, while the skilled group performed significantly better on the NARA-II than the WORD. Therefore, a test with a higher contribution of word decoding to the final score may not be as sensitive to struggling readers as a test with a higher contribution of listening

comprehension skills. Similarly, Nation and Snowling (1997) also compared a group of 17 skilled comprehenders with 17 less skilled comprehenders, ages 7-9 years, on the Suffolk, the NARA-II, the single-word reading subtest of the WORD, the Graded Nonword Reading Test (Snowling, STothard, & Mclean, 1996) and a listening comprehension task. Results indicated that while the skilled group outperformed the less skilled group on single word reading, text level reading, and comprehension, there was not a significant difference between groups on decoding skills. Both of these studies highlight the concern that not all test of reading comprehension will similarly identify students who have deficits in reading comprehension.

Formative Assessments. A further criticism of standardized or norm-referenced tests of reading comprehension is that such tests often have little utility for teachers in instructional planning for individual students, which has lead a number of researchers to investigate the validity of formative assessments, such as curriculum based measurements (CBM), in predicting reading comprehension ability (Fuchs, Fuchs, & Maxwell, 1988). One study focused specifically on the relationship between a CBM for oral reading fluency (ORF) and the subscales of the Woodcock Reading Mastery Test – Revised (Woodcock, 1987), including Passage Comprehension (Hosp & Fuchs, 2005). Using a sample of 310 students in 1st through 4th grade, correlation results indicate a consistently significant and strong relationship between ORF and passage comprehension across all four grades (r=.79-.84). The authors suggest that this finding may be because comprehension requires efficient and accurate reading (Hosp & Fuchs, 2005). Another study examined both ORF and reading comprehension rate, defined as the percentage of comprehension questions answered correctly per minute spent reading, using passages from the Timed Reading series (Spargo, 1989) with the Woodcock Johnson III Tests of Achievement (WJ-III; Woodcock, McGrew, & Mather, 2001) Passage Comprehension subtest and Broad

Reading Composite (Skinner,Williams, Morrow, Hale, Neddenriep, & Hawkins, 2009). The sample for this study consisted of 22 fourth graders and 29 fifth graders from a rural elementary school and 37 tenth graders from an urban high school. Results of a Pearson correlation and standard regression analysis indicated that across all grades, ORF was more strongly correlated to Passage Comprehension than reading comprehension rate, and ORF accounted for 23% of the variance in Passage Comprehension scores for 4th graders, and 35% of the variance for 5th and 10th graders, further reinforcing the strength of ORF as a predictor of reading comprehension.

A third study examined whether four other formative assessments, in addition to ORF, would increase the amount of reading comprehension variance (Marcotte & Hintze, 2009). The other measures included Retell Fluency (RTF), where students orally recount an ORF story and are rated for total words recalled, Sentence Verification Technique (SVT), where students reading four passages and are then asked to indicate which of 16 sentences were included in the passages, Written Retell (WRT), where students read a passage silently for five minutes and then have five minutes to write what they can recall, and a Maze (MZ), or cloze, passage. Formative assessment scores were compared with the Group Reading Assessment and Diagnostic Evaluation (GRADE; Williams, 2001), a norm-referenced assessment of reading ability including a Comprehension Composite, and the Massachusetts Comprehensive Assessment System (MCAS), a state standards based assessment of reading. The sample for this study was a diverse group of 111 4th graders, 80% of whom qualified for free/reduced lunch and 30% identified as English Language Learners. Results of a correlational analysis indicated that all formative assessments including ORF were moderately correlated (r= .46-.67) with GRADE Comprehension Composite scores. Results of a standardized regression indicated that all five formative assessments significantly predicted GRADE Comprehension Composite scores, and

that ORF combined with the other four formative assessments predicted 57% of GRADE variance. A second regression analysis indicated that all formative assessments except RTF significantly predicted MCAS scores and that, combined, the remaining four formative assessments predicted 66% of MCAS variance. These results suggest that ORF have greater validity as a predictor of reading comprehension when considered in conjunction with other formative assessment to give a more complete picture of a reader's skills.

Despite the potential utility of formative assessments in predicting reading comprehension ability described in these studies, it is important to note that some research has questioned the validity of such studies' outcomes for both conceptual and statistical reasons (Paris, Carpenter, Paris, & Hamilton, 2005). Oral reading fluency, in particular, has been conceptually questioned as a correlate of reading comprehension because reading rate may not always be indicative of a student performing both the bottom-up and top-down processes in comprehension, and because reading rate can be influenced by reader characteristics, such as anxiety about reading aloud, and situational factors including text difficulty (Paris et al., 2005). Oral reading fluency as a correlate has also been questioned for statistical reasons, such as a lack of accounting for the floor and ceiling effect often present in reading fluency measures (Paris et al., 2005).

Metacognitive Skill Assessment. Although many measures have clearly addressed cognitive reading comprehension skills, metacognitive skills for reading are generally not explicitly incorporated into reading comprehension assessments (McLain, Gridley, & McIntosh, 1991). The literature on measurement of reading metacognition indicated that most studies investigating the subject used researcher-constructed interviews, but one study revealed a scale of metacognitive skill use during reading which the researchers used when conducting empirical

study of classroom comprehension instruction, the Index of Reading Awareness (IRA; Jacobs & Paris, 1987). The IRA is a multiple choice, self-report scale designed for students in 3rd – 5th grade to rate their evaluation, planning, and regulation abilities during reading. Results of the study indicated the IRA was sensitive to growth in metacognitive skills in an experimental group that received one year of metacognitive strategy instruction (Jacobs & Paris, 1987). However, another study investigating psychometric properties of the scale indicated limited internal reliability (α=0.61) and recommended caution in classroom use of the scale (Mclain, et al., 1991).

Overall, the literature indicates some prominent weaknesses in current reading comprehension assessments. Standardized assessments do not consistently measure the same skills, which can at times be attributed to differences related to item format and item difficulty across different tests. These inconsistencies can lead to variability in which students are identified as struggling with comprehension and also cloud the instructional implications of standardized tests. Some research has suggested that formative tests of reading skills may have utility in predicting reading comprehension ability, but other research contests these assumptions on the basis of both the content of formative assessments and statistical questions. Such findings clearly indicate room for improvement in the area of reading comprehension assessment. Specifically, a need appears to exist for a comprehensive measure of reading comprehension that acknowledges the contribution of metacognitive skills and provides a meaningful link to instructional planning for teachers.

Teacher Judgment

Given the current limitations in reading comprehension assessment instrumentation in psychometric quality, content, and lack of classroom utility, new avenues of assessment

methodology need to be explored. One measurement method that may be appropriate is the use of behavior rating scales. Behavior rating scales provide a format that would allow for psychometrically validated, structured assessments of reading skills that directly utilize teacher judgment. A growing body of research has indicated support for the validity of teacher judgments of reading skills for elementary school students (Hoge & Coladarci, 1989), and recent studies have found moderate to significant correlations between teacher estimates and actual measures of reading outcomes including oral reading fluency measures and reading instructional levels.

Numerous studies have supported teacher's ability to accurately judge the academic achievement of students compared to student performance on standardized achievement tests (Demaray & Elliot, 1998). For example, Bates and Nettelbeck (2001) examined teacher accuracy in predicting the student performance on the reading accuracy and reading comprehension scores of the Neale Analysis of Reading Ability – Revised (NARA-R; Neale, 1988). Twenty-nine general education teachers predicted the percentile rank of 108 6^{th} to 8^{th} grade students for both reading accuracy and reading comprehension. The teachers' estimated percentile rankings were then transformed to NARA-R raw scores by the researchers and these scores were compared with the participating students' actual raw scores on the NARA-R. Results indicated some variability in the accuracy of the teacher ratings, with teachers significantly more likely to over-estimate the reading accuracy and reading comprehension scores of low-achieving students than scores of average and high achieving students. Overall, however correlations between teacher estimates and students achievement in reading accuracy (r= .77) and comprehension (r= .62) fell in the moderate to high range.

Other studies have supported the ability of teachers to accurately judge the reading ability of students as judged by measures of oral reading fluency (Feinberg & Shapiro, 2003). For instance, Madeleine and Wheldall (2005) examined the ability of 33 third through fifth grade teachers to judge the reading ability of twelve students from his or her class by rank ordering their students according to overall reading ability and then comparing these teacher estimations against a measure of oral reading fluency. Students were assessed using five fifth grade-level reading passages from the Wheldall Assessment of Reading Passages (WARP; Wheldall, 1996) that were used to calculate students average words correct per minute (WCPM). Results found that the mean correlation between teacher rankings of student ability and actual student rank as determined by WCPM fell in the high range (r= .73). However, only 5 of teachers accurately identified the three lowest readers identified by WCPM in their classroom, 8 teachers identified at least one student as being in the bottom 25% when in fact the student was in the top 50%, and 10 teachers identified a student as being in the top 25% when in fact the student was in the bottom 50%.

As suggested by the Madeleine and Wheldall (2005) results, some research has indicated that simple correlations might be masking the variability in teacher judgments, and some studies have worked to address this by using statistical procedures such as percentage agreement (Begeny, Krouse, Brown, & Mann, 2011). For example, Begeny, Eckert, Montarello, and Storie (2008) examined the accuracy of teacher judgment in estimating elementary student reading instruction level using a research-generated brief rating scale of reading behaviors called the Teacher Rating Scale of Reading Performance (TRSRP), and estimations of rates of oral reading fluency (ORF). Participants included 10 elementary level teachers and 87 1st to 3rd grade students. Results indicated that, overall, teachers demonstrated acceptably valid judgments about

students' reading ability. Specifically, the researchers found moderate to high levels of correlation between actual and teacher estimated rates of ORF (r= .68) and between TRSRP rating and actual ORF (r= .76), and results indicated 93% accuracy in teacher judgments of which students read at mastery level on grade-level material. Teachers had more difficulty judging students who were at instructional (44% accuracy) and frustrational (42% accuracy) levels of instruction, consistent with other research indicated higher levels of teacher accuracy in predicting the academic performance of high achieving students (Demaray & Elliot, 1998).

Similarly, Begeny and colleagues (2011) also used percentage agreement calculations in their investigation of teacher judgment of reading achievement. In their study, 27 first through fifth grade teachers rated 212 first through fifth grade students (approximately eight students from each teacher's classroom) using the TRSRP, an estimate of the students' words correct per minute (WCPM), an estimate of the students' DIBELS oral reading fluency subtest (DORF; Good & Kaminski, 2002) reading level, and the students' language arts score on the state-wide achievement test, the Palmetto Acheivement Challenge Test (PACT). Students were assessed to determine their actual DORF reading level, WCPM, and PACT language arts proficiency level. Results indicated that the overall correlation between actual and teacher estimated WCPM was r= .51 and the overall correlation between teacher estimated and actual DORF reading level was r= .47. However, a closer examination of the estimated and actual DORF reading levels indicated that while only 55.3% of students accurately identified At Risk and 40.3% accurately identified Some Risk, 71.3% of Low Risk students were accurately identified Low Risk, providing further evidence of teacher's better ability in judging the performance of average and high-achieving students. In addition, among inaccurate DORF reading level estimates, 55.6% were overestimated and 44.4% were underestimated. Results further indicated that correlations

between teacher estimates and actual PACT scores fell in the moderate range r= .58, teachers accurately judged 53.8% of students actual PACT scores, and correlations between actual WCPM and TRSRP scores were also in the moderate range (r= .43).

Feinberg and Shapiro (2009) also addressed statistical ambiguity by using two different correlational methodologies in their study of teacher judgment. The researchers asked 74 second through fifth grade teachers to predict students' reading achievement using the ACES (DiPerna & Elliot, 1999) and to predict student oral reading fluency rates. Teacher predictions were then correlated with actual oral reading fluency scores and with student scores on the Letter-Word Identification and Passage Comprehension subtests of the Woodcock-Johnson Tests of Achievement-III (WIAT; Woodcock, McGrew, & Mather, 2001) using a typical correlation methodology for absolute accuracy and a correlation methodology examining relative accuracy of teacher predictions, corrected for restricted range of estimates. Correlations using the absolute accuracy methodology indicated moderate correlations overall between estimated and actual oral reading fluency scores (r = .64), and between the ACES and Passage Comprehension scores (r = .60), Letter-Word Identification (r = .59) and estimated oral reading fluency rate (r = .47). Correlations using the relative accuracy procedures also found moderate correlations overall between estimated and actual oral reading fluency scores (r = .69), and between the ACES and Passage Comprehension scores (r = .45) and Letter-Word Identification (r = .45). Overall, the results indicated that teachers were significantly better at estimating ORF for average achieving students than low achieving students, and that while they tended to overestimate low achieving ability, there was a substantial range across estimations of low achievement. As in Fienberg and Shapiro's 2003 study, teachers were less accurate at making absolute judgments of student oral reading, but they demonstrated high accuracy in comparing relative rank order of students.

While few studies have explicitly examined the ability of teachers to predict reading comprehension behaviors and outcomes, the significant findings of these studies examining the ability of teachers to predict overall student reading achievement and some measures of reading comprehension suggest that behavior rating scales utilizing teacher judgment may be a valid avenue of reading comprehension assessment warranting further exploration.

Rating Scales of Academic Behaviors

As evident in the teacher judgment literature, a few teacher-completed rating scales of behaviors related to academics do exist and have been utilized in research. For instance, Begeny and colleagues (2008) generated the 9-item TRSRP assessing students' decoding, reading accuracy, reading fluency, reading comprehension, and application of reading skills to school work for their teacher judgment study. Begeny and colleagues (2011) also used the TRSRP for another teacher judgment study. Other teacher judgment studies have highlighted existing, validated, brief rating scales of academic competence such as the Academic Competence scale of the Social Skills Rating System –Teacher (SSRS-T; Gresham & Elliot, 1990), which contains questions pertaining to both reading-specific behaviors and broader academic constructs such as intellectual functioning and motivation (e.g. Demaray & Elliott, 1998).

Outside of the teacher judgment literature, validation studies continue to expand new iterations of existing rating scales of academic behaviors, such as the Academic Competence Evaluation Scales (ACES; DiPerna & Elliot, 1999). In a research brief, Elliott, Huai, and Roach (2007) describe how an experimental instrument call the Brief Academic Competence Evaluation Scales System (BACESS; Elliott, Huai, & DiPerna, 2004) could potentially be used in conjunction with the DIBELS to screen for students with reading problems. The BACESS is based on work done with the ACES and consists of three phases; students are assessed at

subsequent phases when determined to be at risk on a previous phase. In Phase 1, students are rated on a five-level continum for Reading, Language Arts and Mathematics, in Phase 2, students are rated using the Academic Competence scale of the ACES, and in Phase 3 students are rated using the entire ACES. A preliminary validation study of Phase 1 and Phase 2 has indicated some support for this model of student assessment (Kettler & Elliott, 2010). In the validation study, 29 teachers in 2^{nd}-5^{th} grade classrooms rated all students in their classroom on both Phase 1 and 2 of the BACESS, and results were compared to student scores on the Measure of Academic Progress (MAP; South Carolina state standardized achievement test). Results indicated high validity for both Phase 1 (α = .93) and Phase 2 (α = .95). Concurrent validity calculations indicated moderate correlations for Phase1 and Phase 2 with MAP Reading scores (Phase 1 r= .65, Phase 2 r= .50) and MAP Mathematics scores (Phase 1 r= .56, Phase 2 r= .40). Multiple regression results indicated that Phase 1 rates were a significant predictor of future MAP total achievement scores, but Phase 2 scores were not a significant predictor.

Several studies have also begun to explore the utility of rating scales of academic behaviors as screeners for student learning problems (Taylor et al., 2001). One such study examined whether teacher ratings of students' progress in phonics are a valid screener for learning disability in reading (Snowling, Duff, Petrou, & Schiffeldrin, 2011). First grade students (n=146) were assessed for reading problems using the Letter-Sound Knowledge, Word Reading, Sound Deletion, and Sound Isolation subtests from the York Assessment of Reading for Comprehension (YARC; Hulme et al., 2009) along with various cognitive processing screeners. Classroom teachers then rated all participating students using a researcher-generated rating scale that asked for estimates of student progress through the phases of phonics as defined by the London Department for Children, Schools, and Families. Results indicated that students

identified as at-risk on the teacher rating scale scored significantly lower on the YARC than students not identified as at risk and was very effective at identifying students who would go on to have reading difficulty (sensitivity = 0.88) but somewhat less successful at predicting students who would go on to be free of reading disability (specificity = 0.61).

In another study, researchers proposed a model for screening upper elementary students for reading problems using teacher rating scales as component of the universal screening battery (Speece, Ritchey, Sileverman, Schatschneider, Walker, & Andrusik, 2010). Participants included 230 fourth grade students from 20 elementary general education classrooms. Student assessments included the reading comprehension subtest of Gates-MacGinitie Reading Test, the Letter-Word Identification, Word Attack, and Passage Reading Fluency subtests Woodcock-Johnson Tests of Achievement, Third Edition (WJ-III; Woodcock, McGrew, Mather, & Shrank, 2001), the Test of Silent Word Reading Fluency (TOSWRF; Mather, Hammill, Allen, & Roberts, 2004), and the Test of Word Reading Efficiency (TOWRE; Togesen, Wagner, & Rashotte, 1999). In addition, for each child, classroom teachers were asked to complete the Academic Competence scale of the SSRS (Gresham & Elliott, 1990), the Attention Deficit Hyperactivity Disorder Rating Scale – IV (ADHD-IV; DuPaul, Power, Anastapoulos, & Reid, 1998), and the Teacher Reading Rating Form, a researcher generated, teacher rating scale that had items pertaining to word decoding, reading fluency, vocabulary, reading comprehension, and motivation. Researchers found that an efficient, three-factor model (reading comprehension, word fluency, and teacher ratings) for screening of reading problems was identified. Together, reading comprehension, word fluency, and teacher ratings of reading accounted for 46% of variance in discriminating between at-risk and not at-risk students, adding further support to the idea that teacher-completed rating scales of academic skills have utility in the context of screening for learning problems.

One study has also begun to examine the utility of teacher-completed rating scales as an indicator of which students are responding well to intervention and which students are continuing to struggle (Vaughn, Wanzek, Murray, Scammacca, Linan-Thompson, & Woodruff, 2009). Researchers examined the effectiveness of an intensive reading intervention for 1st grade students included a teacher rating of academic competence. Students were assessed at the beginning of 1st grade using the several reading assessments including the DIBELS oral reading fluency subtest (DORF; Good & Kaminski, 2002) and teachers rated students' overall academic competence using the Academic Competence subscale of the SSRS (Gresham & Elliott, 1990). Those at risk were administered 13-26 weeks of intensive reading intervention during 1st grade. At the beginning of 2nd grade, all students who received intervention were re-assessed using DORF. Those who met 2nd grade DORF benchmarks were considered "high-responders" to intervention, while those who continued to fall below benchmark were considered "low-responders." The study results indicated that teachers rated students who fell in the "low-responders" group significantly lower in overall academic competence than students who fell in the "high-responders" group, suggesting that teachers can accurately identify students who are responding adequately to intervention using a rating scale measure.

Evaluation of Academic vs. Behavioral Targets

It is important to note that, while emerging evidence supports the use of rating scales to evaluate academic targets (e.g. reading ability), rating scales have more traditionally been used to evaluate social behavior targets and there are some differences in the nature of behavioral targets relative to academic targets. Specifically, social behavior is understood to be influenced by the context of the behavior (e.g. who the child is interacting with, what the behavioral expectations are in that setting) and therefore ratings of behavior are expected to vary to some degree across

situations, settings, and raters (Ogden, 2003; Elliot, Busse, & Gresham, 1993). Because academic rating scale targets represent more stable constructs (e.g. reading skill rather than academic engagement) that we might not anticipate to exhibit as great of a degree of variability across contexts as we would with social behaviors, we might also expect a lesser degree in variability of ratings across settings and raters.

For example, when examining externalizing behaviors at school, such as symptoms of Attention Deficit Hyperactivity Disorder, it is expected that the intensity and frequency of such behaviors may vary across classroom settings, so it is considered best practice to apply the "aggregate rule" and obtain rating scales from a variety of sources in order to establish a more complete picture of the student's behavior (Merrell, 2000). However, reading ability is unlikely to be influenced as greatly by environmental factors such as classroom behavior, and so it may be expected that, if the RSAS-RCN is truly a measure of reading ability, the RSAS-RCN total scores would demonstrate greater stability across settings than might be expected in scales rating social targets.

Direct Behavior Rating Scales: A Promising Future Direction for the RSAS-RCN

Reading comprehension may also lend itself to observation in the classroom using more direct methodologies than traditional behavior rating scales because many key instructional goals as recommended by the National Reading Panel, including question answering and making predictions, may also be appropriate for progress monitoring (NRP, 2000). Direct behavior ratings evolved from related scales such as Daily Behavior Report Cards and may represent an advantage over traditional, retrospective behavior scales for use in progress monitoring because they rate an individual's behavior immediately after it occurs and are ideal for repeated measurement of a skill over time (Chafouleas, Riley-Tillman, & Christ, 2009).

A growing database of literature has indicated preliminary support for the reliability and validity of DBR single item scales (DBR-SIS). For example, in one study, researchers investigated the use of DBR-SIS for "works to resolve conflict" and "interacts cooperatively with peers" behaviors in a class of 15 preschoolers (Chafouleas, Christ, Riley-Tillman, Briesch, & Chanese, 2007). Over period of 13 days, two teachers and two teacher assistants completed DBR ratings twice per day for all students. Results indicated that when generalizability (G)-studies were conducted within a restricted universe of generalization (within rater, multiple ratings of each student), G- and dependability (D) -coefficients reached the acceptable – highly acceptable range (.87-.93) (Chafouleas et al., 2007). Similarly, a different study examined the use of DBR-SIS for "academic engagement" in a kindergarten classroom with 12 students and 2 teachers. Using an infinite universe of generalization accounting for multiple raters, G- and D-coefficients also reached acceptability (.82 and .77, respectively; Briesch, Chafouleas, & Riley-Tillman, 2010).

In addition to reliability and validity, research has also indicated adequate teacher acceptability ratings of DBR's. For example, Chafouleas, Kilgus, and Hernandez (2009) asked participating teachers in their study to fill out the Assessment Rating Profile – Revised (ARP-R; Eckert, Hintze, & Shapiro, 1997), a scale that asks professionals to rate the acceptability of scales with a series of questions, and using a six-point scale. Results indicated that teachers rated the DBR highly overall (M=4.92) and strongly endorsed a preference for the DBR over the SSRS on a question comparing the two scales. Thus, while clearly still in preliminary stages, data on the reliability, validity, and teacher acceptability of DBR scales so far is encouraging. More research on DBR scales is certainly warranted, but so far, research indicates that DBR is a valid system for evaluating classroom behaviors such as academic engagement and disruptive

behaviors and therefore may be a meaningful avenue to pursue in evaluating academic skill behaviors once valid, traditional rating scale items have been established, particularly as a means of progress monitoring.

Statement of Purpose

Clearly, current reading comprehension measures available are inadequate psychometrically and tools for systematically assessing reading comprehension in the classroom do not exist. A need exists for a reliable, valid screening tool with high levels of feasibility that can be used in the classroom to identify students who are struggling with reading comprehension. Research suggests that teacher judgment of reading skills may be an accurate and valid source of information regarding student progress, and rating scales of academic skills have a brief but promising presence in the literature. This evidence suggests that teacher ratings of reading comprehension may be an untapped resource for assessment. The purpose of this investigation was to address the current limitations of reading comprehension assessments, explore a new methodology for reading comprehension assessment, and expand the current literature on academic rating scales by piloting a validation of a behavior rating scale of reading comprehension for upper elementary students.

CHAPTER III. METHOD

Participants

Teachers and student participants for this study were recruited from eight elementary schools in three Pennsylvania school districts. School districts were recruited through personal and email contact with district and building administrators, and school districts were offered opportunities for free professional development seminars in exchange for participation. Once school district and building administrators agreed to participation in the project, individual third, fourth, and fifth grade teachers were sent a letter of recruitment requesting their participation in the project.

Students were then identified for recruitment in classrooms of teachers who consented to participate. All participating districts had data from DIBELS oral reading fluency (DORF) benchmark assessment already in place as a universal screening measure for reading in their elementary schools. Using percentile rank data from the winter DORF benchmark assessment, all students' scores within each participating teacher's classroom were arranged in a distribution from lowest to highest score. Students with identified reading disabilities were included in an effort to recruit students with a wide range of reading ability levels. Within each classroom, students' were divided into three groups; 1) those at or above the 75th percentile (high performing readers), 2) those between the 25th and 75th percentile (average performing readers), and 3) those below the 25th percentile (low performing readers). One student from within each of these groups was randomly selected for recruitment, so that each teacher would rate one high performing reader, one average performing reader, and one low performing reader. Letters of consent were sent home to the students' parents and those students for whom consent was granted were offered the chance to become participants in the study. Students whose parents

consented were given an assent form, which was read aloud to them. Students who indicated assent by checking a "yes" box on the assent form and signing their name then became participants in the study.

The above recruitment procedures resulted in 41 general education elementary teachers participating in the study, 17 third grade teachers, 14 fourth grade teachers, and 10 fifth grade teachers. Two teachers were male and 39 were female. Five teachers reported a bachelor's degree as their highest educational degree, 25 reported earning a master's degree, and 11 reported a master's degree plus additional graduate credits earned. Reported years of teaching experience ranged from one to 38 years, with an average of 12 years of teaching experience.

Three students were recruited from each of the 41 teachers' classrooms for a total of 123 students recruited. In two classrooms, one of the three parents refused consent, and two additional students from other classrooms refused assent, resulting in a total of 119 student participants. 51 students were in the third grade, 40 in the fourth grade, and 28 in the fifth grade. 55 students were male and 64 were female.

Setting

This study took place in various elementary schools across three school districts in Eastern Pennsylvania. Two districts were large, suburban districts and one was a large rural district. Within the districts, students range from 79-90% white/Caucasian, 11-12% of students are reported to have an Individualized Education Plan (IEP), and 14-33% of students are considered economically disadvantaged as determined by qualification for free or reduced lunch, as per data published by the Pennsylvania Department of Education, 2012 (see Table 1). All participating students were currently enrolled in the 3rd, 4th, and 5th grades.

Measures

Rating Scale of Academic Skills – Reading Comprehension. The Rating Scale of Academic Skills – Reading Comprehension, Narrative (RSAS-RCN) is an individually administered rating scale of specific, elementary school level reading comprehension skills. The RSAS-RCN was conceptualized as a tool for teacher use in the classroom setting to aid in directing instruction, and thus was based on instructional definitions of reading comprehension. The RSAS-RCN was developed as a single-use screening too, with intended long-term potential as a repeated screening tool and future potential for development into a progress-monitoring tool. In order to develop the conceptual basis for this measure, a research team consisting of three graduate research assistants and Dr. Edward S. Shapiro at Lehigh University's Center for Promoting Research to Practice conducted a comprehensive review of the literature pertaining to reading comprehension, with a focus on instructional definitions of reading comprehension. Key sources included the National Reading Panel report *Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction* (2000), information from the University of Oregon's Center on Teaching and Learning (http://ctl.uoregon.edu), Edmonds and Briggs (2003), and Guthrie and Scafiddi (2004), all of which provided detailed information about what teachers are expected to teach children with regard to reading comprehension in the classroom.

The research team reviewed the sources and identified patterns of reading comprehension constructs across the various sources by creating a matrix to group similar reading comprehension skills. Similar constructs were grouped together to develop broad reading comprehension skill domains. Construct groupings were compared and discussed until the team reached a consensus, which resulted in the following five domains: Identifying Information; Understanding Text Structure; Monitoring Comprehension; Retelling/Summarization; and

Analyzing Text. Initial domain definitions and a total of 26 items across the five domains (3-10 items per domain) were also developed based on the information gathered through the construct matrix. Minor refinements of domain definitions and item wordings were made in consultation with a larger research team of Dr. Shapiro, Dr. Mary Beth Calhoon, and six graduate research assistants and this document constituted the initial version of the RSAS-RCN.

Domain Validation. A targeted sample of experts comprised of six university faculty members and consultants for state education agencies was recruited through personal email contact by Dr. Shapiro to serve as an expert panel to validate the initial RSAS-RCN domains and domain definitions. Participants completed a brief electronic rating scale assessing the importance of each domain to the mastery of reading comprehension (4 point scale, 0 = Not At All, 1= Low, 2 = Somewhat, 3 = High) and how well each domain definition matched its corresponding domain (4 point scale, 0 = Not at all, 1 = Weak, 2 = Adequate, 3 = Strong), and were given the opportunity to provide feedback regarding all of the domains and domain definitions (see Appendix A for example items). To make the review of results systematic across items, the research team decided criteria on a cut-off of a mean rating 2.5 or higher on both scales for a domain to be considered "validated."

Results of the expert panel indicated that 4 out of the 5 domains and domain definitions were fully validated. Based on feedback provided by the experts and an overall match rating of 2.33, the domain name "Identifies Information" was changed to "Identifies Content." In addition, although importance and match ratings met the specified criteria for validation, the word "critically" was also removed from the definition for the "Monitoring Comprehension" domain based on noted feedback. No other modifications to the domains or domain definitions were made.

Item Validation. A two-part process was undertaken for initial RSAS-RCN item validation. First, a group of 24 elementary teachers (3 reading specialists and 21 classroom teachers) was recruited from a rural school district to validate the items. The three elementary reading specialists completed a paper-and-pencil rating scale asking participants to select the domain the item belonged to, their degree of certainty about the domain selection, and the importance of the item to the selected domain for all 26 initial RSAS-RCN items. A total of 21 classroom teachers (7 3rd grade teachers, 7 4th grade teachers, and 7 5th grade teachers) completed abbreviated versions of the form, which asked participants to perform the same rating tasks for eight or nine of the initial RSAS-RCN items. The classroom teachers only examined a subgrouping of the total RSAS items in the interest of saving the teachers time and increasing participation in the validation. Thus, each of the 26 items was rated 10 times, 3 times by the reading specialist and 7 times by classroom teachers.

The research team then reviewed the results of the teacher item validation. To ensure systematic review of results across items, the research team decided on a cut-off criteria of at least 7 out of ten teachers identifying that the item belonged in the domain for which it was developed and a mean rating 2.5 or higher on the certainty and importance scales for an item to be considered "validated." If the criteria for the item was not met, the research team examined the responses to see where the ratings fell short and then decided whether to modify the item wording to clarify the item's intended meaning, eliminate the item, or leave the item unchanged to see what the feedback would be during the second phase of item validation. The results of the teacher validation indicated that 9 items were fully validated and 16 were not. For the items that were not fully validated, the research team decided to eliminate three items, change the wording of 11 items, and leave two items unchanged pending feedback from the second phase of item

validation. Corresponding changes were made to all item validation forms, resulting in a total of 23 items to undergo the second stage of validation.

For the second stage of the item validation process, a convenience sample of ten university faculty members and consultants for state education agencies were contacted by Dr. Shapiro through personal email and asked to validate the RSAS-RCN items, as revised by the teacher item validation feedback. For the expert validation surveys, the revised 23 RSAS-RCN items were divided into three sets of 8-9 items and turned into three electronic surveys asking participants to select the domain the item belonged to, their degree of certainty about the domain selection, and the importance of the item to the selected domain (see Appendix B). The purpose of only sending the experts a subset of the items was to reduce the length of the rating task and increase response rate. Each of the experts was emailed one of the surveys containing a subset of the items. Of the ten experts contacted, eight responded, resulting in each item of the RSAS-RCN being rated between 1 and 4 times by experts. The results of the expert item validation were then reviewed by research team. Using the same evaluation criteria as for the teacher item validation (at least 70% of raters identifying the item belonged in the domain for which it was developed and a mean rating 2.5 or higher on the certainty and importance scales for an item to be considered "validated"), the majority of items were fully validated. However, due to the low sample size for some items (8 items were rated by only 1 expert participant due to low response rate on one of the abbreviated surveys), the research team decided not to make changes to any of the items at this stage in the validation process and to instead move forward with the RSAS-RCN items that resulted from the teacher validation, with the anticipation that further changes may result from the empirical item analysis conducted during this pilot study.

The resulting pilot RSAS-RCN scale is a 23-item rating scale that asks teachers to estimate how regularly their students engage in specific behaviors related to reading comprehension of narrative text using a 7-point rating scale (0=Never, 6=Always; see Appendix C for sample items), a scaling gradient that has been found to have acceptable levels of rater variance in past studies (Chafouleas, Christ, & Riley-Tillman, 2009). A total score was then calculated based on the summation of ratings across individual items.

Group Reading Assessment and Diagnostic Evaluation. The Group Reading Assessment and Diagnostic Evaluation (GRADE; Williams, 2001) is a group-administered series of norm-referenced tests of reading ability. The GRADE assesses 11 areas of reading skills from preschool through early college, including phonological awareness and simple word understanding in prekindergarten and sentence comprehension, passage comprehension, and word reading beginning in 1st grade. Subtests assessing different skills are administered at developmentally grade-appropriate times; each year in school is administered a different level of the GRADE. The GRADE provides standard scores, percentile ranks, normal curve equivalents, stanines, grade equivalents, and a Growth Scale Value (for progress monitoring purposes) for individual subtests, as well as for a total test score.

The GRADE is reported to have strong psychometric properties (Fugate, 2001). Internal reliability for the total test scores across all grade levels ranges from .81 to .94, while the reading comprehension subtest internal reliability is reported as .94. Test-retest reliability coefficients across grades range from .77 to .98. Content validity between the GRADE Total Reading score and the Iowa Test of Basic Skills ranges between .69 and .83 for 4th and 5th grades, and between the Gates-MacGinitie Reading Test ranges from .86 to .90 across 1st, 2nd, 3rd, and 6th grades. Predictive validity of the GRADE to the TerraNova ranges from .76 to .86 across 2nd, 4th, and 6th

grades (Fugate, 2001). For this study, only the two subtests comprising the Reading Comprehension Composite (Sentence Comprehension and Passage Comprehension) were administered, and the Reading Comprehension Composite standard score was used for purposes of analysis.

Pennsylvania System of School Assessment - Reading. The Pennsylvania System of School Assessment (PSSA) is a group-administered, standards-based, criterion reference measure that examines a student's attainment of state-determined academic standards. All students in Grades 3-11 are assessed annually in reading and math. The PSSA Reading test assesses students in two main areas; 1) comprehension and reading skills, and 2) analysis and interpretation of fictional and non-fictional text. The Reading test requires students to read grade-appropriate passages and respond to both multiple choice and open-ended questions. Individual student scores are reported as scaled scores and classified as "Below Basic," "Basic," "Proficient," or "Advanced." Classifications are based on different cut point scores, which vary across each subject for each grade. For example, in 3rd grade, PSSA Reading scaled scores below 1168 are considered "Below Basic," scaled scores between 1168 to 1235 are considered "Basic," scaled scores from 1235 to 1442 are considered "Proficient," and scaled scores above 1442 are considered "Advanced." Reliability coefficients alphas for 3rd, 4th, and 5th grades ranged from .74 to .91 for Reading (Data Recognition Corporation, 2011). For the purposes of analysis in this study, scaled scores were utilized.

Social Validity Scale. The researchers also developed a brief acceptability questionnaire examining the social validity of the RSAS-RCN (see Appendix D) that all participating teachers had the opportunity to complete. The questionnaire included questions addressing how long teachers took to administer the RSAS-RCN, how many students they could feasibly complete an

RSAS-RCN on at one time, and space for any additional comments. The data from this scale included simple descriptive statistics (e.g. range and mean) and qualitative comments.

Procedures

This study took place during the month of May and the first week of June. As detailed in the Participants section, recruited teachers were given letters of explanation and consent for participation in the study, and letters of explanation and consent were then sent home to parents of identified students in consenting teachers' classrooms. Students were also asked to provide written assent to participate in the study. All consenting students were taken out of class for approximately 50 minutes to be group-administered the reading comprehension composite portion of the GRADE.

All participating teachers met briefly with a research assistant to review the RSAS-RCN and administration procedures. After looking over the RSAS-RCN and answering any questions the teachers had about the form, teachers were instructed that, for the consenting students in their classroom, they were to informally observe the students reading comprehension behaviors during class time with the RSAS-RCN items in mind for two weeks. The purpose of this observation period was to help teachers anchor their ratings to current observations of student behavior and to standardize the methodology for administering the measure across teachers. At the end of the two-week period, they were asked to complete an RSAS-RCN for each consenting student. Teachers were given the choice to complete a paper-and-pencil version of the scale or an electronic version of the scale. Two teachers chose to complete the rating scales electronically, and 39 teachers chose the paper version. Teachers were also asked to respond to a brief acceptability survey after completing the rating scales. A total of 38 teachers completed the acceptability survey. To examine test-retest reliability, a group of 6 randomly selected teachers

were asked to re-administer the RSAS-RCN a second time to the same three students in their classroom, following the same procedures as the first administration. The second RSAS-RCN was completed one week after the first administration of the RSAS-RCN. All six teachers chose to use the paper-and-pencil version of the scale for the retest sample. Finally, researchers requested the release of PSSA Reading score data for consenting students after students completed the test at their school during the spring semester.

Throughout the data collection process, measures were also taken to ensure the confidentiality of the participants. The CBM data used to identify student participants within classrooms and the PSSA test data were collected electronically through the use of databases provided by the school district. This method of data collection eliminated the need to access individual student files in order to collect data. In addition, the principal investigator worked with the principals and reading specialists at participating schools to ensure that only the necessary standardized test score report information were viewed and recorded. The principal investigator was the only data collector in this study and was the only individual who administered and recorded the GRADE to participating students. Finally, teachers were asked to identify students on the RSAS-RCN forms that they completed only by the students' existing student identification number (previously assigned by the school districts). Once all data were collected by the principal investigator, all teacher and student names were immediately removed and only identification numbers were used during data analysis.

Data Analyses

Traditionally, psychoeducational scales have been developed and validated using classical test theory (CTT), which is predicated on test-level information and the assumptions of a true score, uncorrelated error scores, and a linear relationship between true, error, and observed

scores (Novick, 1966). CTT item statistics are sample dependent and test reliability and construct validity under CTT are often examined using procedures including Cronbach's alpha and factor analysis (Lei, Wu, DiPerna, & Morgan, 2009). CTT has the advantages of fairly simple analytic procedures, relatively low sample size requirements, and a long history of use in the psychometric literature; however, CTT also has disadvantages, most importantly the fact that all person and item statistics are sample-dependent which can lead to low item discrimination and reliability and validity estimates that are influenced by the variability of test respondents' abilities (Lei et al., 2009).

Another test development theory, known as item response theory (IRT), offers a methodology that compensates for some of the weaknesses inherent in CTT. Unlike CTT, IRT is based on item-level (rather than test-level) information and IRT models are probabilistic and non-linear (rather than linear) in nature. This allows for calculation of item and person statistics that are sample independent and the calculation of item characteristics and ability scores, which may be considered advantages of IRT over CTT (Hambleton & Jones, 1993).

Given the strengths and weaknesses of both CTT and IRT, this study utilized statistical procedures including traditional (CTT typical) analyses including correlational procedures to examine test-retest reliability, and single-parameter, IRT-based Rasch modeling procedure (Rasch, 1960) to investigate construct validity. When Rasch models were originally developed, they were intended for only dichotomous data. Since then, a family of Rasch models have been developed to address expanded testing situations including polytomous data (Bond & Fox, 2007). Because the RSAS-RCN data was polytomous, an Andrich Rating-Scale Model version of Rasch modeling, which allows for polytomies, was utilized for the construct validity analysis. The complete data analysis for this project followed a three step process.

Pre-Analysis. First, a series of power analyses for correlations were conducted with an $\alpha = .05$, for a medium effect size, and statistical power of .80, determined that a minimum of 85 participants were needed for determining significant differences between correlations (Cohen, 1992; Mayr, Erdfelder, Buchner, & Faul, 2007). Guidelines for constructing polytomous Rasch models indicate that each rating category should contain a minimum of 10 observations, which could be met with as few as 70 participants, although a sample size of 100 participants is recommended for robust Rasch findings (Green & Frampton, 2002; Linacre, 2004). Finally, the recommended sample size for exploratory factor analysis is between 5 and 10 subjects per scale item; as the RSAS-RCN has 23 items, between 115 and 230 subjects were indicated for exploratory factor analysis (Costello & Osborne, 2005). The pool of 119 student participants for this study met these parameters.

Preliminary Analyses. Data screening procedures were followed to detect any problems with the data set. Frequency distributions of all RSAS-RCN items scores, RSAS-RCN items scores, the RSAS-RCN total score, GRADE Reading Comprehension Composite standard scores, and PSSA-Reading scaled scores were examined to determine ranges of scores and check for the presence of outliers. A Mahalanobis distance procedure was also conducted to investigate the presence of any outliers. Frequencies of item scores were then plotted and analyses of item skewness and kurtosis were determined for RSAS-RCN items scores, the RSAS-RCN total score, GRADE Reading Comprehension Composite standard scores, and PSSA-Reading scaled scores to examine normality. To further examine whether data were appropriate for correlational analyses, Q-Q plots of RSAS-RCN total scores, GRADE Reading Comprehension Composite standard scores, and PSSA-Reading scaled scores were compiled to check for homoscedasticity (homogeneity of variance) and bivariate scatterplots between RSAS-RCN initial and retest total

scores, between RSAS-RCN initial total scores and GRADE Reading Comprehension Composite standard scores, and between RSAS-RCN initial total scores and PSSA-Reading scaled scores were examined for linear covariance and presence of outliers. Finally, a Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity were conducted to determine if the RSAS-RCN initial total score data met adequate criteria (KMO of 0.5 or greater and significance on Bartlett's test of sphericity) for exploratory factor analysis. A brief examination of the means of the three primary measures, specifically the RSAS-RCN initial total scores, the GRADE Reading Comprehension Composite scores, and the PSSA-Reading scaled scores, across participants grouped by winter ORF scores (students at or below the 25th percentile, students between the 25th and 75th percentile, and students at or above the 75th percentile) was calculated to determine if the measures differentiated students of varied reading abilities.

Analysis of Hypotheses. Analyses to address the projects' specific research questions were conducted.

Research Question 1: Is there adequate evidence for the construct validity of the RSAS-RCN, including substantive validity and structural validity? Construct validity was primarily assessed using Rasch modeling. However, because the applicability of Rasch modeling is dependent on the assumption of unidimensionality, or that the items included define a single dominant construct, an exploratory factor analysis was first conducted to identify latent constructs (dimensions) in the RSAS-RCN (Green & Frampton, 2002). EFA factor structures were evaluated using standard multiple criteria including eigenvalues greater than 1.0, Catell's (1966) scree test, at least 10% of variance explained for each identified factor, internal consistency (Cronbach's alpha) of .40 or higher, minimization of items loading on multiple

factors, and degree to which the factor solution is theoretically meaningful (Costello & Osborne, 2005; McDermott, 1993).

Once the RSAS-RCN construct structure was identified through EFA, RSAS-RCN item- and person-level data was entered into the Winsteps software program (Linacre, 2007) and a Rasch analysis was conducted to further investigate the item structure and explore how the items perform in relation to students with differing levels of reading comprehension ability. First, the model's unidimensionality was confirmed by examining the model's explained and unexplained variance. Next, category function statistics were examined to assess the structural validity of the RSAS-RCN's 7-point scale. Specifically, the function of the assessment was deemed acceptable if each category count was 10 or greater, average measures and step measures ($\pm$ SE) were ordered, the mean-square Outfit statistic for each category was less than two, and the category probability map indicated that categories were evenly distributed and each was likely to be endorsed by a proportion of respondents (Linacre, 2002). Next, substantive validity was assessed in the following three ways. First, mean square infit and outfit values were reviewed to examine each item's conformity to the overall model; consistent with criteria for a Likert-type rating scale, values between 0.6 – 1.4 were considered acceptable (Wright & Linacre, 1994). Next, the person separation reliability coefficient (considered conceptually equivalent to Cronbach's alpha) and the item reliability coefficient (indication of how well items are spread along the construct continuum) were evaluated; values above .70 were considered sufficiently reliable (Bond & Fox, 2007). Finally, the person-item map was examined to determine the range in individual students' ability and item difficulty and to examine the degree of overlap between person ability and item difficulty plots (Bond & Fox, 2007). Adjustments and further iterations

of the Rasch model were conducted as indicated to determine the optimal format for the RSAS-RCN.

Research Question 2: What is the test-retest reliability of the RSAS-RCN total score? Pearson Product Moment correlations were conducted between the two administrations of the RSAS-RCN for the sub-sample of students who were assessed twice to determine test-retest reliability. Consistent with typical interpretations of correlations found in the literature, a correlation $\geq$ .70 was considered strong, a correlation between .40 and .69 was considered moderate, and a correlation of .39 or lower was considered weak (Evans, 1996).

Research Question 3: What is the external validity of the RSAS-RCN with a standardized assessment of reading comprehension (e.g. GRADE)? Pearson Product Moment Correlations were used to determine concurrent validity between RSAS-RCN total scores and the GRADE Reading Comprehension Composite scores, following the same guidelines for correlation interpretation described for Research Question 2. In addition, Pearson Product-Moment correlations were also run between the RSAS-RCN total scores and the GRADE Reading Comprehension Composite scores across students grouped by ability level as determined by winter ORF scores to determine if the relationship between the RSAS-RCN and GRADE remained consistent across reading levels.

Research Question 4: What is the diagnostic validity of the RSAS-RCN to levels of reading proficiency as determined by a state-wide, standards-based assessment of reading (e.g. PSSA-Reading)? Diagnostic validity was first examined by conducting a Pearson Product Moment Correlation between RSAS-RCN total scores and the PSSA-Reading scaled scores, again following the same guidelines for correlation interpretation described for Research Question 2. Similarly to the external validity analysis procedures, Pearson Product-Moment

correlations were also run between the RSAS-RCN total scores and the PSSA-Reading scaled scores across students grouped by reading ability levels to determine if the relationship between the RSAS-RCN and PSSA-Reading remained consistent across reading levels.

Diagnostic validity was also assessed using a receiver operating characteristic (ROC) curve analysis of the RSAS-RCN total scores to PSSA-Reading scores re-classified as either "proficient" or "not proficient" using grade-specific cut-off scores, per PA state guidelines. ROC curve analysis is a statistical procedure that generates an index of the sensitivity (or proportion of true positives to identified positives) against one minus the specificity (or proportion of true negatives to identified negatives) for a scale to a binary classifier system (McFall & Treat, 1999). In this case, the scale is the RSAS-RCN total score, and the binary classifier system is the PSSA-Reading "proficient" or "not proficient" score. ROC curve analysis results in both a graphical plot of the curve above the line of chance and a statistic known as the area under the curve (AUC), which is equal to the probability that a classifier (e.g. the RSAS-RCN) will rank a randomly chosen positive instance higher than a randomly chosen negative one (e.g. will accurately discriminate between "proficient" and "not proficient" on the PSSA-Reading). Scales that result in an AUC of 0.80-0.90 are considered good classifiers and those at 0.90 or higher are considered excellent classifiers (Distefano & Morgan, 2011).

Research Question 5: What is the social validity of the RSAS-RCN to teachers, as measured by an informal, teacher-completed questionnaire pertaining to acceptability? The social validity of the RSAS-RCN was examined through the use of an informal, researcher-generated acceptability questionnaire (see Appendix B). A total of 38 teachers completed the survey (15 third grade teachers, 13 fourth grade teachers, and 10 fifth grade teachers), and 3

declined to complete it. The data from this scale was examined using simple descriptive statistics (e.g. range and mean) and comments were qualitatively summarized.

CHAPTER IV. RESULTS

The purpose of this study was to conduct a pilot psychometric validation of the RSAS-RCN. All data were examined using multiple procedures including preliminary data checks, correlational analyses, factor analysis, Rasch modeling, and cross-tabulation procedures to determine multiple facets of RSAS-RCN scale reliability and validity.

Preliminary Analyses

Frequency distributions did not indicate missing data for any RSAS-RCN items or the RSAS-RCN total score. Two missing cases were identified for GRADE Reading Comprehension Composite standard scores, and both were for students who were absent during the administration period and could not be rescheduled before the end of the school year. Four missing cases were identified for PSSA scores, and all four were students unintentionally left out of PSSA score data provided by one school district. Several attempts were made to obtain the missing cases, but the schools did not respond. Students missing GRADE or PSSA reading data were excluded from the correlational and cross-tabulation analyses, per conventional guidelines (Leong & Austin, 2006). Frequency distributions of all RSAS-RCN items scores, the RSAS-RCN total scores, GRADE Reading Comprehension Composite standard scores, and the PSSA-Reading scaled scores also did not indicate the presence of any outliers or incorrectly entered data. A screening of the data using a Mahalanobis distance procedure did not indicate the presence of any multivariate outliers. Results of the analyses of item skewness and kurtosis for RSAS-RCN items scores, the RSAS-RCN total scores, the GRADE Reading Comprehension Composite standard scores, and the PSSA-Reading scaled scores indicated that all scales fell within acceptable limits (-2 to 2) of skewness and kurtosis (Leong & Austin, 2006). Visual examination of frequency plots of item scores for the same items confirmed the normality of

these data. Q-Q plots of RSAS-RCN total scores and GRADE Reading Comprehension Composite standard scores run to check for homoscedasticity (homogeneity of variance), an assumption for correlational analyses, were also visually examined and found to display acceptable degree of homoscedasticity for all data. Bivariate scatterplots between RSAS-RCN initial and retest total scores and between RSAS-RCN initial total scores and GRADE Reading Comprehension Composite standard scores examined for bivariate normality indicated linear relationships between the two sets of variables and did not indicate the presence of any outliers that could influence correlation statistics. Finally, a Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity indicated that the data met adequate criteria for exploratory factor analysis.

An examination of the descriptive statistics of the RSAS-RCN total scores, GRADE Reading Comprehension Composite Scores, and PSSA-Reading scaled scores across reading ability groups, as identified by winter ORF scores originally used to nominate students for participation in the study, was also conducted. A examination of the means and standard deviations of RSAS-RCN total scores found distinctly different averages for students with ORF scores at or below the 25th percentile (M=71.08, SD=23.26), students with ORF scores between the 25th and 75th percentile (M=92.02, SD=20.142), and students with ORF scores at or above the 75th percentile (M=115.07, SD=17.39; see Figure 1). Further, a one-way analysis of variance indicated significant differences in RSAS-RCN total scores across reading ability groups, $F(2, 116) = 46.304$, $p<.01$. Similar stepwise patterns across reading ability groups were also found for the means of GRADE Reading Comprehension Composite Scores, and PSSA-Reading scaled scores (see Figures 2 and 3).

Research Question 1: Is there adequate evidence for the construct validity of the RSAS-RCN, including substantive validity and structural validity?

In order to determine a possible structure for the 23 items in the RSAS-RCN, an exploratory factor analysis of the measure was conducted. Eigenvalues, percentage of explained variance, and a scree plot of the possible factor solutions were used to make initial decisions for further factor analysis. Using the criterion of eigenvalue larger than 1, eigenvalues indicated a single factor solution. Examination of the percent of variance explained (using a criterion of ideal variance ≥ 5%) also suggested that a one-factor solution was most appropriate; the first factor explained 78.2% of variance, and all following factors accounted individually for less than 3% of the remaining variance. The scree plot illustrated a pronounced drop after the first factor, further confirming that a one-factor solution was indicated. Thus, only a one-factor model for the data was analyzed. Using a .4 loading criterion, Factor 1 retained all 23 RSAS-RCN items (see Table 2) and had a reliability coefficient ($\alpha = .98$) that met desired criterion (.7 or higher).

Given that the EFA supported a unidimensional model of the RSAS-RCN items, data for all 23 items were entered into WINSTEPS (Linarce, 2007) and a Rasch model analysis of the RSAS-RCN was run. First, the scale's unidimensionality was confirmed by examining the table of standardized residual variance. Results indicated that the 23 RSAS-RCN items conformed to a single dimension (conceptually analogous to the single factor described by the EFA), with 78.8% of raw variance explained by the Rasch measure (Eigenvalue unit = 85.4).

Next the assessment function (structural validity) of the RSAS-RCN's 7-point scale format of response was evaluated using category structure data (see Table 3). All seven categories met the criteria of at least 10 observations, average measures and step measures were ordered (e.g. average measures increased incrementally from a value of -3.60 for category 0 to a

value of 6.65 for category 6; step measures increased incrementally from a value of -6.41 for category 1 to a value of 6.07 for category 6). Six of the seven categories met criteria (value less than 2) for the outfit meansquare; category 0 outfit meansquare was 2.72, which indicates this category may be problematic. Finally, a plot of category probabilities indicated that the categories were evenly distributed and each is likely to be endorsed by a proportion of respondents (see Figure 4).

The item infit mean-square and outfit mean-square statistics were then examined. All items fell within acceptable range for outfit mean-square scores (RSAS-RCN items range 0.69 to 1.39), but for infit mean-square scores, one item (Item #8, Domain 5) fell outside the desired range (infit mean-square = 1.43), and another item (Item #4, Domain 2) fell just within the desired range, suggesting that these items may be problematic. All other items fell within the desired range of infit mean-square scores (remaining RSAS-RCN items range 0.69 to 1.18). Item and person separation indices were then examined to determine the reliability of the scale; both measures indicated high levels of reliability (person reliability = .99, item reliability = .97). Finally, the distribution of students and items on the person-item map (see Figure 5) indicated a broad range of student ability (e.g. items were clearly easy for above average participants) and narrower range of item difficulty, with all items demonstrating overlap with student in the mid- to lower-end of the map. This suggests that the RSAS-RCN is very effective at discriminating low-performing comprehenders and less effective at discriminating skill levels amongst above average-performing comprehenders, which is consistent with its intended use as a screener for identifying students struggling with reading comprehension.

Because of the indicated problems with category 0 and two of the RSAS-RCN items discussed above, the Rasch analysis was rerun to see if removing Item #8, Domain 5 and Item

#4, Domain 2 and collapsing scale categories 0 and 1 into a single category would optimize the scale model output. First, the modified scale's unidimensionality was checked. Results indicated that the 21 RSAS-RCN items still conformed to a single dimension, with 79.6% of raw variance explained by the measure (Eigenvalue unit = 82.2) with only 2.5% of raw variance unexplained (Eigenvalue unit = 2.6).

Next, the assessment function (structural validity) of the modified RSAS-RCN's 6-point scale format of response was evaluated using category structure data (see Table 4). All six categories met the criteria of at least 10 observations, average measures and step measures were ordered (e.g. average measures increased incrementally from a value of -4.92 for category 1 to a value of 5.59 for category 6; step measures increased incrementally from a value of -4.33 for category 2 to a value of 5.0 for category 6). All six fell well below 2 for the outfit meansquare, indicating that collapsing categories 0 and 1 solved the outfit problem from the initial model. Finally, a plot of category probabilities indicated that the categories were evenly distributed and each is likely to be endorsed by a proportion of respondents (see Figure 6).

Item infit meansquare and outfit meansquare statistics were examined. All 21 items fell within acceptable range for outfit meansquare scores (range 0.70 to 1.30) and infit meansquare scores (range 0.71 to 1.27). Item and person separation indices were then examined to determine the reliability of the scale; both measures indicated high levels of reliability (person reliability = .98, item reliability = .97). The distribution of students and items on the person-item map (see Figure 7) again indicated a broad range of student ability and narrower range of item difficulty, with all items demonstrating overlap with student in the mid- to lower-end of the map.

Research Question 2: What is the test-retest reliability of the RSAS-RCN total score?

A Pearson Product-Moment correlation was computed between the first and second administration of the 23-item RSAS-RCN total scores for the subset of student participants who were randomly selected for the retest group (n=18) to determine test-retest reliability. Results indicated a strong relationship between the first and second administration total scores, $r(16)=0.95$, $p<.01$.

Research Question 3: What is the external validity of the RSAS-RCN with a standardized assessment of reading comprehension (e.g. GRADE)?

External validity of the 23-item RSAS-RCN to a standardized assessment of reading comprehension was assessed by computing a Pearson Product-Moment correlation between the RSAS-RCN total scores and the GRADE Reading Comprehension Composite standard scores. Results indicated a moderate relationship between the RSAS-RCN and the GRADE Reading Comprehension Composite, $r(115)=0.66$, $p<.01$.

Pearson Product-Moment correlations were also conducted between the RSAS-RCN and the GRADE Reading Comprehension Composite across students grouped by ability level as determined by winter ORF scores. Results indicated a significant, moderate relationship between the RSAS-RCN total score and the GRADE Reading Comprehension for students with ORF scores at the 25th percentile or lower, $r(34)=0.45$, $p<.01$, and for students with ORF scores between the 25th and 75th percentile, $r(40)=0.40$, $p<.01$. Results indicated a non-significant, low relationship between the RSAS-RCN total score and the GRADE Reading Comprehension Composite for students with ORF scores at or above the 75th percentile, $r(37)=0.23$, $p=.17$.

Research Question 4: What is the diagnostic validity of the RSAS-RCN to levels of reading proficiency as determined by a state-wide, standards-based assessment of reading (e.g. PSSA-Reading)?

Diagnostic validity of the RSAS-RCN total score was first examined by conducting a Pearson Product-Moment correlation between the RSAS-RCN total scores and PSSA-Reading scaled scores. Results indicated a relationship at the strong end of the moderate range between the RSAS-RCN total score and the PSSA-Reading scaled score, $r(113)=0.68$, $p<.01$. Similar Pearson Product-Moment correlations were also conducted between the same scores across students grouped by ability level as determined by winter ORF scores. Results indicated that for students with ORF scores at the 25th percentile or lower, there was a non-significant, low relationship between the RSAS-RCN total score and the PSSA-Reading scaled score, $r(32)=0.31$, $p=.07$. For students with ORF scores between the 25th and 75th percentile, results indicated a significant, moderate relationship between the RSAS-RCN total score and the PSSA-Reading scaled score, $r(39)=0.46$, $p<.01$, and for students with ORF scores at or above the 75th percentile, results also indicated a significant, moderate relationship between the RSAS-RCN total score and the PSSA-Reading scaled score, $r(38)=0.60$, $p<.01$.

Diagnostic validity of the RSAS-RCN to the PSSA-Reading was further assessed using a ROC curve analysis. PSSA-Reading scaled scores were re-coded into a binary classification of "proficient" or "not proficient" based on grade-specific cut-off scores, and RSAS-RCN total scores were kept in their original form. Results of the ROC analysis found a high AUC estimate of .841, $p \leq .001$, 95% CI = .77 - .92. In other words, ROC curve analysis indicated an 84% likelihood that students with a Proficient classification on the PSSA will have a higher score on the RSAS-RCN that those student who are not Proficient. These results were confirmed by examining a plot of the ROC curve above a reference line indicating chance, or AUC of 0.5 (see Figure 8).

Research Question 5: What is the social validity of the RSAS-RCN to teachers, as measured by an informal, teacher-completed questionnaire pertaining to acceptability?

A total of 38 teachers completed the RSAS-RCN Acceptability Survey. Overall, teachers reported that they spent between 5 and 15 minutes to complete the RSAS-RCN for one student. When asked how many scales they believed they could feasibly complete at a given time (e.g. as a screening tool), teachers responded with numbers ranging from one to twenty, with a median response of five scales at a time. When asked if they believed they could repeatedly administer the RSAS-RCN for specific students (e.g., as a progress monitoring tool), 32 teachers responded "yes" and six teachers responded "no." The survey also provided teachers space to give any additional comments, and 13 teachers utilized this portion of the survey. A number of comments indicated teachers' perceptions of the RSAS-RCN's strengths, such as "I think the domains and definitions are thorough and make the teacher reflect and answer purposefully," and "I like that it assessed a broad range of skills." Other comments addressed concerns related to the observational nature of the scale (e.g. "I felt like I was only guessing for many questions," and "Some of the questions were difficult, due to the fact that they were personal strategies the student would use independently,") and the utility of the full scale as a potential progress-monitoring tool (e.g. "My concern is whether I would have enough 'evidence' to support my ratings of my students if done repeatedly").

CHAPTER V. DISCUSSION

Construct Validity Results

In general, the results of both the exploratory factor analysis and the Rasch analysis of the data supported the RSAS-RCN as a highly reliable, single-construct scale. Results of the exploratory factor analysis provided preliminary evidence of the RSAS-RCN's high internal reliability, as captured by a high alpha statistic, and indicated that a one-factor model best fit the data. Rasch modeling confirmed the single-factor structure of the scale and the high internal reliability, as represented by high person and item reliability statistics. Visual examination of the person-item plot indicated a broad range of student ability, and a restricted range of item difficulty, which is consistent with what might be expected for a criterion-referenced scale, or any scale used to assess whether students have mastered a specific set of skills. An examination of category data indicated that all criteria were met for 6 of the 7 categories of the 7-point scale used to rate RSAS-RCN items, suggesting that the RSAS-RCN could be improved by collapsing two point categories and changing the scale from a 0-6 to a 1-6 point rating system. Examination of item-level output indicated that 21 out of 23 items met criteria for full inclusion, suggesting that the RSAS-RCN might also be improved by eliminating the two problematic items.

Based on these findings, a second Rasch analysis was conducted on a revised 21-item, 6-point scale model of the RSAS-RCN. Results of this secondary Rasch analysis indicated that the model retained its high reliability, single-construct nature, and once again visual examination of the person-item map indicated a pattern consistent with desired distribution of item difficulty for a screening tool. In this secondary model, all 6 categories and all 21 items also met full inclusion criteria, suggesting that a revised version of the RSAS-RCN may indeed provide even stronger data for use in the classroom setting.

Test-Retest, External, and Diagnostic Validity Results

Overall, results of this study indicated that the RSAS-RCN yielded consistent results over repeated administrations and a moderate to strong relationship with other tests of reading abilities. The strong relationship of the test-retest correlation ($r(16)=0.95$, $p<.01$) indicated a very similar performance between RSAS-RCN total scores from the first and second administration of the RSAS-RCN and suggested high test-retest reliability. The correlation between RSAS-RCN total scores and GRADE Reading Comprehension Composite scores fell at the high end of the moderate range ($r(115)=0.66$, $p<.01$), indicating acceptably strong external validity. Prior research examining teacher estimates of overall reading achievement and more specific reading skills including oral reading fluency found similar results of correlations in the high moderate range. This study's results are consistent with those findings and extend the literature to demonstrate that teachers are equally able to estimate students' reading comprehension skills as other reading skills.

An examination of the relationship between the RSAS-RCN and the GRADE across students grouped by reading ability level indicated similarly moderate relationships for students in the low and average reading ability groups, but a non-significant, low relationship for the high reading ability group. This may suggest that the RSAS-RCN total score does not capture the same range of skills among high readers that the GRADE Reading Comprehension composite score captures, but it must be noted that the sample size for these correlations was very low and further investigation of these relationships with a larger sample size is warranted.

Correlational analysis of the relationship between RSAS-RCN total scores and PSSA-Reading scaled scores also found a result at the high end of the moderate range. These results are consistent with past studies' findings of teacher judgments of overall academic achievement

relative to standardized measures of achievement fell at the high end of the moderate range (e.g. Hoge & Coladarci, 1989). Further examination of the relationship between the RSAS-RCN and the PSSA-Reading scaled scores across students grouped by reading ability level indicated moderate relationships for students in the average and high reading ability groups, and a non-significant, low relationship for students in the low reading ability group. This may suggest that the RSAS-RCN total score does not capture the same range of skills among struggling readers as the PSSA-Reading scaled score, but again, the possible effect of low sample size for these analyses must be recognized. Such a small sample size likely introduces issues of power into the strength of the correlational results.

The diagnostic validity of the RSAS-RCN total score to the PSSA-Reading test was additionally examined using ROC curve analysis. Results of the ROC curve analysis indicated that the RSAS-RCN total score effectively discriminated student classified as "proficient" vs. "not proficient" on the PSSA-Reading, as indicated by the AUC statistic which met criteria for a "good" classifier (DiStefano & Morgan, 2011). Also, it should be noted that ROC analysis typically requires very large samples to attain stability in findings; the fact that the measure was able to find these AUC values with the size here is as an indication that the diagnostic validity of the measure is likely to hold up with larger samples. The combined evidence of the correlational and ROC curve analyses of the RSAS-RCN and PSSA-Reading data suggests that the RSAS-RCN may be a useful diagnostic tool for identifying students who will and will not be proficient on a state standardized test of reading achievement.

Social Validity Results

Overall, the results of the teacher-completed acceptability survey indicated that most teachers found the 23-item RSAS-RCN to be a potentially useful tool for use in the classroom

setting. A majority of the teachers reported that the RSAS-RCN could be completed in five to 15 minutes per students and that they could administer the measure as a screening tool to at least 5 students at a time. A majority of teachers also indicated that they believed they could repeatedly administer the RSAS-RCN to individual students for use as a repeated-use screening tool (e.g., during multiple benchmark periods throughout the school year). Additional comments by teachers indicated an endorsement of the item's content, and although several teachers expressed doubts related to the validity of their judgment on some items, this study's psychometric investigation of the scale indicates that the teachers were more accurate in their judgments than they might have expected.

Limitations

In interpreting the results of this study, it is also important to acknowledge its limitations. First, there are specific limitations associated with the sample of students and teachers used for this study. For instance, the population for this study was comprised of a sample of teacher and student participants from a limited region of Pennsylvania, and thus the results cannot be generalized to other areas of Pennsylvania or the country at large. Similarly, the school districts included in the study were fairly homogenous in SES, racial and ethnic makeup, and setting (rural and suburban). The results are only reflective of this one, fairly homogeneous, sample. The very small size of the sample also restricts the generalizability of this study's results; although it is encouraging that significance was found for many of the statistical tests run in this study, larger sample sizes with greater variability of students and teachers may yield very different results. The study was also limited to students in upper elementary school grades (3^{rd}-5^{th}); these results cannot be extended to younger or older students, although the scale itself may be applicable to older or younger students' reading comprehension skills. In addition, this study's

sample did not include a proportion of English language learning (ELL) students. Evidence suggests teachers' may be less accurate in discerning and reporting reading difficulties among ELL students than among monolingual English students (Limbos & Geva, 2001), and this study's sample does not allow for a comparison across ELL and English monolingual student groups.

Second, there are also several limitations inherent in the design of the study. Specifically, the data for this study was all collected at the end of the spring term, after the teachers had worked with their students for almost an entire school year. Thus, this data does not generalize to what might be found at other time periods during the same year, even for the same sample. Also, as acknowledged in the method section, this study's design requires all teachers to rate three students. Although this is an effective sampling method to help determine whether the RSAS-RCN differentiated across students with differing levels of reading ability, the low sample size prevented the researchers from accounting for effects of data nesting (students within teachers) in the correlational analyses and eliminated the using of regression in investigating diagnostic validity to the PSSA-Reading.

Finally, there are additional limitations associated with the scope of the scale itself. In particular, the RSAS-RCN specifically addresses students' reading comprehension skills within the context of reading narrative text, excluding an examination of students' skills reading expository text. Research suggests that students' comprehension skills and ability to answer comprehension questions can vary across text types due to the influence of factors including background knowledge and higher order cognitive skills such as planning and organizing (Eason, Goldberg, Young, Geist, & Cutting, 2012). Given these differences, the RSAS-RCN, with its items targeting narrative reading comprehension skills, may not adequately capture the scope of

a students strengths and difficulties in reading comprehension when reading more complex, expository text, and further investigation is warranted.

Implications for Practice

Notwithstanding these limitations, it should be also noted that the results of this pilot study have implications beyond simply the direct feedback regarding the reliability and validity of the RSAS-RCN as a scale. As a behavior rating scale, the RSAS-RCN is a tool that quantifies teacher judgments about students academic skills, and the results of this study suggest that teacher perceptions about upper elementary students' reading comprehension skills are quite accurate. Despite a consistent demonstration of research demonstrating the value of teacher judgments of academic abilities, particularly in the area of reading achievement (e.g. Demaray & Elliot, 1998), beliefs that teacher judgments are too biased to be accurate or a useful source of information in schools has persisted (e.g., Hoge & Coladarci, 1989). Up to this point, teacher judgments about academic rather than social behaviors have been chronically under-utilized in the schools, and there have not been any tools available to teachers that quantify their judgments about specific, standards-derived academic skills that could be regularly used in a screening capacity in the classroom.

The strong validation data provided by this preliminary study of the RSAS-RCN suggests that teacher rating scales of academic behaviors could be a valuable addition to schools' current processes of identifying students with reading problems. Such teacher-completed scales could have a multitude of uses in a school setting, including use as a component of screening processes in a manner consistent with recommended multi-gated procedures for screening for social behavior problems in schools (Severson, Walker, Hope-Dolittel, Kratochwill, & Gresham, 2007).

Hopefully, further research will confirm the utility of the RSAS-RCN and allow it to serve as new, valuable source of data in the school setting.

Future Research Directions

There are a number of important directions future explorations of the RSAS-RCN can take. Foremost, as this study was only a pilot validation of the RSAS-RCN, a large-scale validation study with a much larger sample size should be conducted next to confirm the results of this pilot. Future samples should include a wider diversity of students and teachers from varied SES and racial backgrounds and from various regions of the country. Explicit comparisons between ELL and English monoligual students' ratings should also be investigated to examine possible differences in teacher judgments across these populations. Samples should also be sufficiently large to allow for statistical procedures that would account for the effects of nested data from the results.

It should also be noted that this study's Rasch modeling of the current RSAS-RCN indicated that the scale could be improved by eliminating two items and collapsing a category of the rating scale, resulting in a revised 21-item, 6-point rating scale version of the RSAS-RCN. Future studies may continue to investigate item-level data analyses to confirm whether such results are replicated and use those results to revise the RSAS-RCN to make it as efficient and effective as possible for classroom use. Item-level analyses might also be used in investigating the utility of the scale with broader populations; in particular, item-level analyses might be helpful in targeting which items from the scale are most functional to different grade levels and students with different language backgrounds and use these to develop differentiated versions of the scale as appropriate. As discussed above, future studies may also wish to consider the addition of items or the use of a separate scale to examine students' reading comprehension skills

with expository text, in addition to the narrative text skills already addressed by the RSAS-RCN, to see if such items add to the scale's utility.

Finally, the purpose of this pilot study was to examine the efficacy of the RSAS-RCN as a screening tool, implying that a limited number of administrations would be anticipated within a single school year. However, during the development process of the RSAS-RCN data, a great deal of attention was paid to the literature on direct behavior rating single item scales (e.g., Chafouleas et al., 2009), and RSAS-RCN items were written to reflect highly specific, observable skills that might then be used as a stand alone, progress monitoring measure to assess whether a student is demonstrating growth in a particular area of reading comprehension. Current research suggests that single item scales asking teachers to rate behaviors such as "academic engagement" and "disruptive behavior" can be used repeatedly over time to track a student's progress or can be used in a single administration as a reliable screening tool once an optimal cut-score has been determined (Chafouleas et al., 2007; Kilgus, Chafouleas, Riley-Tillman, & Welsh, 2012). Although the high internal reliability and the clustered nature of items on the Rasch model's person-item maps suggests that all items are so closely related that they may not statistically differentiate across separate reading comprehension skills, it bears investigating whether individual items or the RSAS-RCN as a whole might be effective for use as a progress monitoring tool or an abbreviated screening tool.

Conclusion

The purpose of this study was to conduct a pilot psychometric investigation of the RSAS-RCN and address questions of construct, test-retest, external, diagnostic, and social validity of the scale. The results of the study indicate that the RSAS-RCN functions as a highly reliable, single-construct screening tool, with high test-retest reliability and moderate external, diagnostic,

and social validity. Although further investigation with a larger and more diverse sample is indicated, this study provides strong preliminary evidence that the RSAS-RCN may serve as a useful measure that will fill noted weaknesses both in the reading comprehension scale literature and in practice as a screening tool for students struggling with reading comprehension difficulties in the classroom. This study also makes an important addition to the literature on the accuracy and validity of teacher judgments of students' academic skills in the area of reading comprehension and speaks to the potential utility of ratings scales as a valid method for evaluating academic targets.

References

Bates, C., & Nettelbeck, T. (2001). Primary school teachers' judgments of reading achievement. *Educational Psychology, 21,* 177- 187.

Begeny, J. C., Krouse, H. E., Brown, K. G., & Mann, C. M. (2011). Teacher judgments of students' reading abilities across a continuum of rating methods and achievement measures. *School Psychology Review, 41,* 23-38.

Begeny, J. C., Eckert, T. L., Montarello, S. A., & Storie, M. S. (2008). Teachers' perceptions of students' reading abilities: An examination of the relationship between teachers' judgments and students' performance across a continuum of rating methods. *School Psychology Quarterly, 23,* 43-55. doi: 101037/1045-3830.23.1.43

Bond, T. G. & Fox, C. M. (2007). Applying the Rasch model: Fundamental measurement in the human sciences, 2nd edition. New York, NY: Routledge Taylor & Francis Group.

Bowyer-Crane, C., & Snowling, M. J. (2005). Assessing children's inference generation: What do tests of reading comprehension measure? *British Journal of Educational Psychology, 75,* 189-201.

Briesch, A. M., Chafouleas, S. M., & Riley-Tillman, T. C. (2010). Generalizability and dependability of behavior assessment methods to estimate academic engagement: A comparison of systematic direct observation and direct behavior rating. *School Psychology Review, 39,* 408-421.

Cain, K., & Oakhill, J. (2007). Reading comprehension difficulties: Correlates, causes, and consequences. In Cain, K., & Oakhill, J. (Eds). *Children's Comprehension Problems in Oral and Written Language: A Cognitive Perspective.* New York: Guilford Press.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology, 96,* 31-42. doi: 10.1037/0022-0663.96.1.31

Cantrell, S. C., Almasi, J. F., Carter, J. C., Rintamaa, M., & Madden, A. (2010). The impact of a strategy-based intervention on the comprehension and strategy use of struggling adolescent readers. *Journal of Educational Psychology, 102,* 257-280. doi: 10.1037/a0018212

Chafouleas, S. M., Christ, T. J., & Riley-Tillman, T. C. (2009). Generalizability of scaling gradients on direct behavior ratings. *Educational and Psychological Measurement, 69,* 157-173. doi: 10.1177/0013164408322005

Chafouleas, S. M., Briesch, A. M., Riley-Tillman, T. C., Christ, T. J., Black, A. C., and Kilgus, S. P. (2010). An investigation of the generalizability and dependability of Direct Behavior Rating Single Item Scales (DBR-SIS) to measure academic engagement and disruptive behavior of middle school students. *Journal of School Psychology, 48,* 219-246. doi: 10.1016/j.jsp.2010.02.001

Chafouleas, S. M., Christ, T. J., Riley-Tillman, T. C., Briesch, A. M., & Chanese, J. A. M. (2007). Generalizability and dependability of Direct Behavior Ratings to Assess Social Behavior of Preschoolers. *School Psychology Review, 36,* 63-79.

Chafouleas, S. M., Kilgus, S. P., & Hernandez, P. (2009). Using Direct Behavior Rating (DBR) to screen for school social risk: A preliminary comparison of methods in a kindergarten sample. *Assessment for Effective Intervention, 34,* 214-223. doi: 10/1177/1534508409333547

Chafouleas, S. M., Riley-Tillman, T. C., & Christ, T. J. (2009). Direct Behavior Rating (DBR): An emerging method for assessing social behavior within a tiered intervention system. *Assessment for Effective Intervention, 34,* 195-200. doi: 10.1177/1534508409340391

Christ, T. J., Riley-Tillman, T. C., & Chafouleas, S. M. (2009). Foundation for the development and use of Direct Behavior Rating (DBR) to assess and evaluate student behavior. *Assessment for Effective Intervention, 34,* 201-213. doi: 10.1177/1534508403949390

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

Connecticut Department of Education. (2000). *Connecticut Master Test-Language Arts Handbook.* Retrieved from the World Wide Web: http:www.state.ct.us/sde/

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, & Evaluation, 10,* 1-9.

Cross, D. R., & Paris, S. G. (1987). Assessment of reading comprehension: Matching test purposes and test properties. *Educational Psychologist, 22,* 313-322.

CTB/McGraw-Hill. (1982). *The Comprehensive Test of Basic Skills.* Monterey, CA: Author.

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.

Data Recognition Coporation. (2011). *Technical Report for the 2010 Pennsylvania System of School Assessment.* Maplegrove, MN: DRC. Downloaded on March 19, 2011 from http://www.portal.state.pa.us/portal/server.pt/community/technical_analysis/7447

Demaray, M. K., & Elliot, S. N. (1998). Teacher judgments of students' academic functioning: A comparison of actual and predicted performance. *School Psychology Quarterly, 13,* 8-24.

DiPerna, J. C., & Elliot, S. N. (1999). Development and validation of the academic competence evaluation scales. *Journal of Psycholeducational Assessment, 17,* 207-225.

DiStefano, C. & Morgan, G. (2011). Examining classification crieteria: A comparison of three cut score methods. *Psychological Assessment, 23,* 354-363. DOI: 10.1037/a0021745

Dunn, L. M. & Markwardt, F. C. (1970). *Examiner's manual; Peabody Individual Achievement Test.* Circle Pines, MN: American Guidance Service.

DuPaul, G. J., Rapport, M. D., & Perriello, L. M. (1991). Teacher ratings of academic skills: The development of the Academic Performance Rating Scale. *School Psychology Review, 22,* 284-300.

Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012). Reader-text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology, 104,* 515-528.

Eckert, T. L., Dunn, E. K., Codding, R. S., Begeny, J. C., & Kleinman, A. E. (2006). Assessment of mathematics and reading performance: An examination of the correspondence between direct assessment of student performance and teacher report. *Psychology in the Schools, 43,* 247-265.

Eckert, T. L., Hintze, J. M., & Shapiro, E. S. (1997). School psychologists' acceptability of behavioral and traditional assessment procedures for externalizing problem behaviors. *School Psychology Quarterly, 12,* 150-169.

Edmonds, M., & Briggs, K. L. (2003). The instructional content emphasis instrument: Observations of reading instruction. In S. Vaughn and K. L. Briggs (Eds.). *Reading in the classroom: Systems for the observation of teaching & learning.* Baltimore, MD: Paul H. Brookes.

Elliot, S. N., Busse, R. T., & Gresham, F. M. (1993). Behavior rating scales: Issues of use and development. School Psychology Review, 22, 313-321.

Elliot, 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.

Feinberg, A. B. & Shapiro, E. S. (2009). Teacher accuracy: An examination of teacher-based judgments of students' reading with differing achievement levels. *The Journal of Educational Research, 102,* 453-462.

Feinberg, A. B. & Shapiro, E. S. (2003). Accuracy of teacher judgments in predicting oral reading fluency. *School Psychology Quarterly, 18,* 52-65.

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

Fugate, M. H. (2001). Review of the Group Reading Assessment and Diagnostic Evaluation. In *The Fourteenth Mental Measurements Yearbook.* Available from http://search.ebscohost.com

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

Garner, R. & Kraus, C. (1981-1982). Good and poor comprehender differences in knowing and regulating reading behaviors. *Educational Research Quarterly, 6,* 5-12.

Glover, T. A., & Albers C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45,* 117-135. doi: 10.1016/j.jsp.2006.05.005

Green, K. E., & Frantom, C. G. (2002). Survey development and validation with the Rasch model. Paper presented at the International Conference on Questionnaire Development, Evaluation, and Testing, Charleston, SC.

Gresham, F. M. (2007). Response to intervention and emotional and behavioral disorders: Best practices in assessment for intervention. *Assessment for Effective Intervention, 32,* 214-222. doi: 10.1177/15345084070320040301

Gresham, F. M. & Elliott, S. N. (1990). *Social Skills Rating System.* Circle Pines, MN: American Guidance Service.

Guthrie, J. T., & Scafiddi, N. T. (2004). Reading comprehension for information text: Theoretical meanings, developmental patterns, and benchmarks for instruction. In John T. Guthrie, A. Wigfield, & K. C. Perencevich (Eds.). *Motivating reading comprehension: Concept-oriented reading instruction*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hagley, F. (1987). *Suffolk Reading Scale.* Windsor: NFER-Nelson.

Hacker, D. J. (2004). Self-regulated comprehension during normal reading. In Ruddell, R. B., & Unrau, N. J. (Eds.). *Theoretical Models and Processes of Reading, Fifth Edition,* (pp.755-779. Newark: International Reading Association.

Hambleton, R. K. & Jones, R. W. (1993). An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: Issues and Practice, 38-47.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of the literature. *Review of Educational Research, 59,* 297-313.

Hoover, H., Hieronymus, A., Frisbie, D., & Dunbar, S. (1993). *Iowa Test of Basic Skills.* Chicago: Riverside.

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

Jacobs, J. E., & Paris, S. G. (1987). Children's metacognition about reading: Issues in definition, measurement, and instruction. *Educational Psychologist, 22,* 255-278.

Johnston, A. M., Barnes, M. A., & Desrochers, A. (2008). Reading comprehension: Developmental processes, individual differences, and interventions. *Canadian Psychology, 49,* 125-132. doi: 10.1037/0708-5591.49.2.125

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.

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

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12,* 281-300.

Kettler, R. J., & Elliott, S. N. (2010). A brief broadband system for screening children at risk for academic difficulties and poor achievement test performance: Validity evidence and applications to practice. *Journal of Applied School Psychology, 26,* 282-307. doi: 10.1080/15377903.2010.518584

Kilgus, S.P., Chafouleas, S.M., Riley-Tillman, T.C., & Welsh, M.E. (2012). Direct behavior rating scales as screeners: A preliminary investigation of diagnostic accuracyin elementary school. *School Psychology Quarterly, 27,* 41-50. doi: 10.1037/a0027150

Kintsch, W. (2004). The construction-integration model of text comprehension and its implications. In Ruddel, R. B. & Unrau, N. J. (Eds.). *Theoretical Models and Processes of Reading, Fifth Edition* (pp. 1270-1328). Newark: International Reading Association.

Kintsch, W. & Kintsch, E. (2005). Comprehension. In Paris, S. G. & Stahl, S. A. (Eds.), *Children's Reading Comprehension and Assessment* (pp. 3-12). New York: Routledge.

Lei, P., Wu, Q., DiPerna, J. C., & Morgan, P. L. (2009). Developing short forms of the EARLI numeracy measures: Comparison of item selection methods. Educational and Psychological Measurement, 69, 825-842. doi: 10.1117/0013164409332215

Leong, F. T L., & Austin, J. T. (2006). *The psychology research handbook: A guide for graduate students and research assistants.* Thousand Oaks: Sage Publications.

Leslie, L. & Caldwell, J. (2001). *Qualitative Reading Inventory-3*. New York: Addison Wesley Longman.

Limbos, M. M., & Geva, E. (2001). Accuracy of teacher assessments of second-language students at risk for reading disability. *Journal of Learning Disabilities, 34,* 136-151.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3,* 85-106.

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith Jr & R. M. Smith (Eds.), Introduction to Rasch measurement, (pp.258-278). Maple Grove, MN: JAM Press.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). *Gates-MacGinitie Reading Tests* (4th ed.). Itasca, IL: Riverside.

Madelaine, A., & Wheldall, K. (2005). Identifying low-progress readers: Comparing teacher judgment with a curriculum-based measurement procedure. *International Journal of Disability, Development and Education, 52,* 33-42.

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. doi: 10.1016/j.jsp.2009.04.003

McBride, J. R., Ysseldyke, J., Milone, M., & Stickney, E. (2010). Technical adequacy and cost benefit of four measures of early literacy. *Canadian Journal of School Psychology, 25,* 189-204. doi: 10.1177/082957357351036796

McDermott, P. A. (1993). National standardization of a uniform multisituational measure of child and adolescent psychopathology. *Psychological Assessment, 5,* 413-424.

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50,* 215-41.

McLain, V. A. M., Gridley, B. E., & McIntosh, D. (1991). Value of a scale used to measure metacognitive reading awareness. *Journal of Educational Research, 85,* 81-88.

Merrell, K. W. (2000). Informant report: Rating scale measures. In Shapiro, E. S., & Kratochwill, T. R. (Eds.). *Conducting School-Based Assessments of Child and Adolescent Behavior.* New York: The Guilford Press.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Nation, K., & Snowling, M. (1997). Assessing reading difficulties: The validity and utility of current measures of reading skill. *British Journal of Educational Psychology, 67,* 359-370.

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

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

Neale, M. D. (1989). *The Neale Analysis of Reading Ability-Revised.* Windsor: NFER.Oakhill, J. V., Cain, K., & Bryant, P. E. (2002). The dissociation of word reading and text comprehension: Evidence from component skills. *Language and Cognitive Processes, 18* 443-468.

Novick, M.R. (1966) The axioms and principal results of classical test theory. Journal of Mathematical Psychology, 3, 1-18.

Ogden, T. The validity of teacher ratings of adolescents' social skills. Scandinavian Journal of Educational Research, 47, 63-76. doi: 10.1080/003138032000033335

Paris, S. G., Carpenter, R. D., Paris, A. H., & Hamilton, E. E. (2005). Spurious and genuine correlates of children's reading comprehension. In Paris, S. G. & Stahl, S. A. (Eds.), *Children's Reading Comprehension and Assessment* (pp. 3-12). New York: Routledge.

Pearson, D. P. & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices – past, present, and future. In Paris, S. G. & Stahl, S. A. (Eds.), *Children's Reading Comprehension and Assessment* (pp. 3-12). New York: Routledge.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests.*Copenhagen: Danish Institute for Educational Research (Expanded edition, 1980. Chicago: University of Chicago Press).

Riedel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly, 42,* 546-567. doi: 10.1598/RRQ.42.4.5

Riley-Tillman, T. C., Chafouleas, S. M., Briesch, A. M., & Eckert, T. L. (2008). Daily behavior report cards and systematic direct observation: An investigation of the acceptability, reported training and use, and decision reliability among school psychologists. *Journal of Behavioral Education, 17,* 313-327. doi: 10.1007/s10864-008-9070-5.

Riley-Tillman, T. C., Chafouleas, S. M., Sassu, K. A., Chanese, J. A. M., & Glazer, A. D. (2008). Examining the agreement of direct behavior ratings and systematic direct observation data for on-task and disruptive behavior. *Journal of Positive Behavior Interventions, 10,* 136-143. doi: 10.1177/1098300707312542

Skinner, C. H., Williams, J. L., Morrow, J. A., Hale, A. D., Neddenriep, C. E., & Hawkins, R. O. (2009). The validity of reading comprehension rate: Reading speed, comprehension, and comprehension rates. *Psychology in the Schools, 46,* 1036-1048. doi: 10.1002/pits. 20442.

Smith, E.V., Conrad, K.M., Chang, K., & Piazza, J. (2002). An introduction to Rasch measurement for scale development and person assessment. Journal of Nursing Measurement, 10, 189-206.

Snowling, M. J., Duff, F., Petrou, A., & Schiffeldrin, J. (2011). Identification of children at risk of dyslexia: the validity of teacher judgments using "Phonic Phases." *Journal of Research in Reading, 34,* 157-170. doi: 10.111/j.1467-9817.2011.01492.x

Spear-Swerling, L. (2004). Fourth graders' performance on a state-mandated assessment involving two different measures of reading comprehension. *Reading Psychology, 25,* 121-148. doi: 10.1080/027027104904

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.

Sweet, A. P. (2005). Assessment of reading comprehension: The RAND Reading Study vision. In Paris, S. G. & Stahl, S. A. (Eds.), *Children's Reading Comprehension and Assessment* (pp. 3-12). New York: Routledge.

Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology, 38,* 934-947.

Taylor, H. G., Anselmo, M., Foreman, A. L., Schatschneider, C., & Angelopoulos, J. (2000). Utility of kindergarten teacher judgments in identifying early learning problems. *Journal of Learning Disabilities, 33,* 200-210.

Vaughn, S., Wanzek, J., Murray, C. S., Scammacca, N., Linan-Thompson, S., & Woodruff, A. L. (2009). Response to early reading intervention: Examining higher and lower responders. *Exceptional Children, 75,* 165-183.

Wechsler, D. L. (1992). *Wechsler Individual Achievement Test*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1990). *Wechsler Objective Reading Dimensions.* London: Psychological Press.

Wiederholt, J. L., & Bryant, B. R. (2001). *Gray Oral Reading Test-IV.* Austin, TX: Pro-Ed.

Wiederholt, J. L., & Bryant, B. R. (1992). *Gray Oral Reading Test-III.* Austin, TX: Pro-Ed.

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

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement.* Itasca, IL: Riverside.

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

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8,* 370.

Appendix A.

Direct Academic Ratings: Reading Comprehension – Narrative Version Domain Validation

DIRECTIONS

The following is a list of five reading comprehension domains which encompass the skills needed in third through fifth grades for mastery of reading comprehension. Each domain includes a definition which attempts to accurately match the domain listed. For this task, consider narrative text. Please do the following for each domain:

A. **Rate how important each domain is to the mastery of reading comprehension:**
N=Not at all L = Low S = Somewhat H = High

B. **Rate how well each definition matches its corresponding domain:**
0=Not at all 1 = Weak 2 = Adequate 3 = Strong

C. **Provide any additional comments or suggestions in the space following each domain's row.**

Domain	Importance N=Not at all L=Low S=Somewhat H =High	Definition	Match 0=Not at all 1=Weak 2=Adequate 3=Strong
Identifies Information	N L S H	Student answers or generates questions of varying levels of complexity about narrative text.	0 1 2 3
Suggestions:			

Domain	Importance N=Not at all L=Low S=Somewhat H =High	Definition	Match 0=Not at all 1=Weak 2=Adequate 3=Strong
Understands Text Structure	N L S H	Student accurately uses text structure to help recall text content and aid understanding in narrative text.	0 1 2 3
Suggestions:			

© Edward Shapiro, 2012

Appendix B.

Direct Academic Ratings: Reading Comprehension – Narrative Version Content Validation

DIRECTIONS

You are being asked to organize reading comprehension skills into corresponding domains. Below is a list of five reading comprehension domains which encompass the skills needed in grades three through five for later mastery of reading comprehension. Next to each domain there is a definition which represents the domain listed. Please review each of the seven statements on the following page and indicate: ***the domain to which each statement corresponds***, ***the certainty of your domain choice***, ***and how important you believe the statement is to the mastery of the domain you selected***. Please begin by familiarizing yourself with the domains and their definitions. You may remove this page for your reference in order to complete the rating tasks.

Domains	Definitions
I. Identifies Content	Student answers or generates questions of varying levels of complexity about narrative text.
II. Understands Text Structure	Student accurately uses text structure to help recall text content and aid understanding in narrative text.
III. Monitors Comprehension	Student demonstrates awareness of his/her own level of understanding of the text and changes reading behaviors accordingly.
IV. Retells/Summarizes	Student directly restates or synthesizes information from the text in his/her own words.
V. Analyzes Text	Student examines narrative text to make predictions, connections and inferences.

RATING TASKS

Please be certain to evaluate each statement by providing the following 3 ratings:

A. Indicate the domain to which each statement best corresponds by circling the appropriate numeral.
I – Identifies Content **II** – Understands Text Structure **III** – Monitors Comprehension
IV – Retells/Summarizes **V** – Analyzes Text

B. Indicate the ***certainty of your domain choice*** for each item by circling:
1 = Not very certain **2** = Relatively certain **3** = Very certain

C. Indicate how ***important*** you believe the ***mastery of this item is to the domain*** you selected by circling:
L = Low/Not important **S** = Somewhat important **H** = Highly important

© Edward Shapiro, 2012

	Narrative Text Statement	Domain					Certainty			Importance		
1	Student accurately identifies similarities and makes cross-text comparisons across narrative text selections.	I	II	III	IV	V	1	2	3	L	S	H
2	Student accurately summarizes narrative text (e.g., prioritizes, chunks, or synthesizes and expresses information).	I	II	III	IV	V	1	2	3	L	S	H
3	Student accurately identifies the main idea in narrative text.	I	II	III	IV	V	1	2	3	L	S	H
4	Student accurately uses multiple points of view to analyze the main idea of narrative text.	I	II	III	IV	V	1	2	3	L	S	H
5	Student accurately determines what he/she does not understand (e.g., identifies where they are having difficulties) in narrative text.	I	II	III	IV	V	1	2	3	L	S	H
6	Student accurately makes predictions based on content in narrative text.	I	II	III	IV	V	1	2	3	L	S	H
7	Student accurately answers identifies information by answering literal questions (i.e., answers that are directly stated in the text) about narrative text.	I	II	III	IV	V	1	2	3	L	S	H

Thank you for volunteering your time to help in the creation of this *DAR Reading Comprehension – Narrative Version Measure*.

© Edward Shapiro, 2012

Appendix C.

Rating Scale of Academic Skills: Reading Comprehension – Narrative

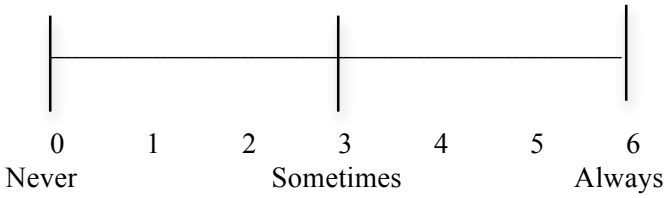
DIRECTIONS

You are being asked to rate the accuracy with which your student performs skills in five domains of reading comprehension. Please review each of the 23 statements and estimate ***how often your student accurately performs each skill***.

RATING TASKS

Please be certain to evaluate **each** statement by providing the following rating:

A. Rate your student's overall performance in each domain, based on the domain's given definition.
B. Rate how regularly your student accurately performs each skill within the domains by ***circling the appropriate numeral next to each item***.



Domain I	Definition	Rating
Identifies Content	Student answers or generates questions of varying levels of complexity about narrative text.	0 1 2 3 4 5 6

Domain I	Definition	R
Identifies Content	Student answers or generates questions of varying levels of complexity about narrative text.	
Item	**Skill**	**R**
1	Student accurately identifies information by answering evaluative questions (i.e., answers that require response based on opinion, beliefs, or personal experiences) about narrative text.	0 1 2
2	Student accurately answers identifies information by answering literal questions (i.e., answers that are directly stated in the text) about narrative text.	0 1 2
3	Student accurately identifies information by answering inferential questions (i.e., answers that are indirectly stated in the text) about narrative text.	0 1 2

Appendix D.

RSAS-RCN Teacher Acceptability Survey

Name: _____

Grade currently teaching:_____

Approximately how many minutes did it take you to fill out the RSAS-RCN? _____

If the RSAS-RCN were to be used as a screening measure, how many RSAS-RCN's do you think you could feasibly complete at a time?

_____

Do you think you could repeatedly administer the RSAS-RCN on specific students (circle one)?

Yes No

Any additional comments:

_____

_____

_____

Table 1

*Summary of Demographic Data for Participating School Districts**

District (Percentage of sample)	Descriptive Designation	Enrollment by Race/Ethnicity	Percent with IEP's	Percent Economically Disadvantaged**
District A (35%)	Large, Suburban	90% White/Cacucasian 4% Hispanic 3% Asian 2% African American	12%	14%
District B (21%)	Large, Rural	88% White/Caucasian 4% Hispanic 3% Asian 3% African American	11%	33%
District C (44%)	Large, Suburban	79% White/Caucasian 4% Hispanic 3% Asian 2% African American 2% Multi-ethnic	12%	14%

* District-wide demographic data as report by the Pennsylvania Department of Education, 2012
** As determined by qualification for free or reduced lunch

Table 2

Eigenvalues and Commonalities for RSAS-RCN from SPSS Factor Analysis Program

Variable	Communality	Factor	Eigenvalue	Percentage of Variance	Cumulative Percentage of Var.
Item # 1	.761	1	17.984	78.19	78.19
Item # 2	.735	2	0.66	2.87	81.06
Item # 3	.834	3	0.55	2.37	83.45
Item # 4	.737				
Item # 5	.730				
Item # 6	.716				
Item # 7	.683				
Item # 8	.764				
Item # 9	.810				
Item # 10	.830				
Item # 11	.874				
Item # 12	.780				
Item # 13	.779				
Item # 14	.795				
Item # 15	.804				
Item # 16	.816				
Item # 17	.821				
Item # 18	.789				
Item # 19	.767				
Item # 20	.834				
Item # 21	.807				
Item # 22	.710				
Item # 23	.807				

Table 3

Summary of Category Structure for full 23-item, 7-point scale RSAS-RCN

Category Label	Observed Count	Observed Average*	Structure (Step) Measure	Structure (Step) S.E.	Outfit Meansquare
0	10	-3.60	None	-----	2.72
1	123	-3.16	-6.41	0.36	0.81
2	244	-1.16	-2.77	0.13	1.07
3	506	0.37	-1.15	0.09	0.98
4	726	2.28	0.99	0.07	0.89
5	701	4.27	3.27	0.06	0.93
6	427	6.65	6.07	0.08	1.04

*Describe as "average measures" in text.

Table 4

Summary of Category Structure for modified 21-item, 6-point scale RSAS-RCN

Category Label	Observed Count	Observed Average	Structure (Step) Measure	Structure (Step) S.E.	Outfit Meansquare
1	115	-4.92	None	-----	0.79
2	221	-2.55	-4.33	0.14	1.14
3	458	-0.93	-2.52	0.09	1.03
4	659	1.07	-0.26	0.07	0.91
5	646	3.18	2.11	0.07	0.92
6	400	5.59	5.0	0.09	1.05

Figure 1. RSAS-RCN total score means by reading ability group.

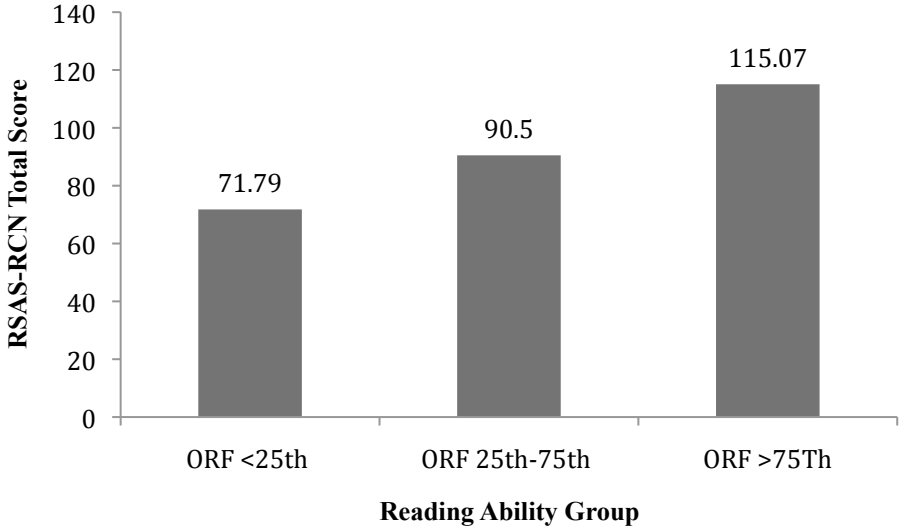


Figure 2. GRADE Reading Comprehension Composite score means by reading ability group.

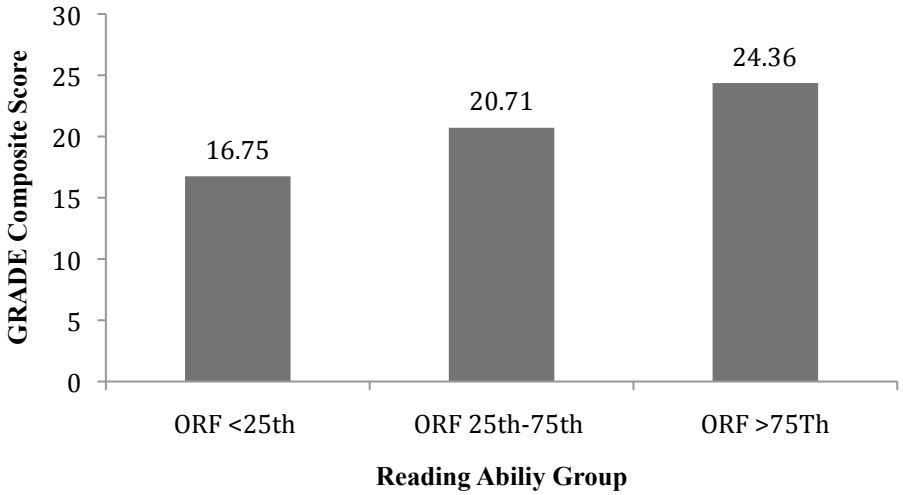


Figure 3. PSSA-Reading scaled score means by reading ability group.

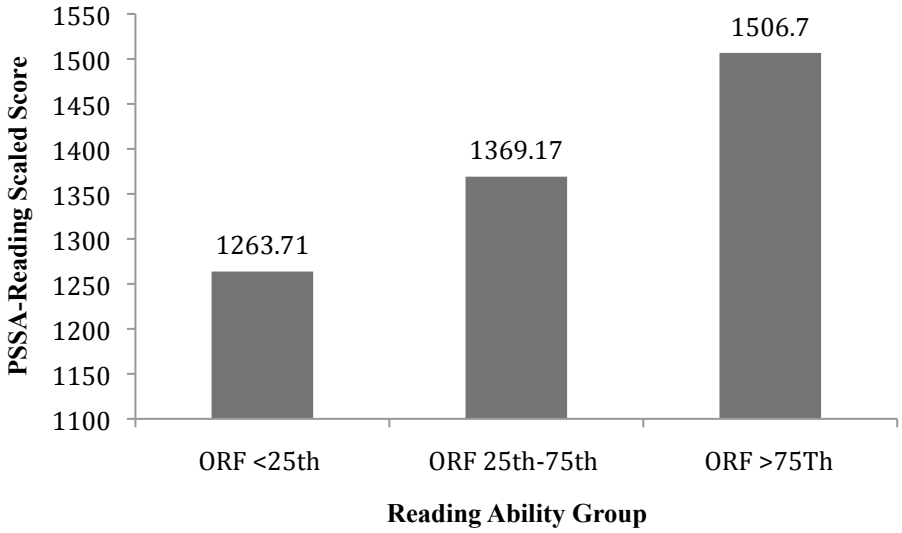


Figure 4. Category probabilities for full 23-item, 7-point scale RSAS-RCN.

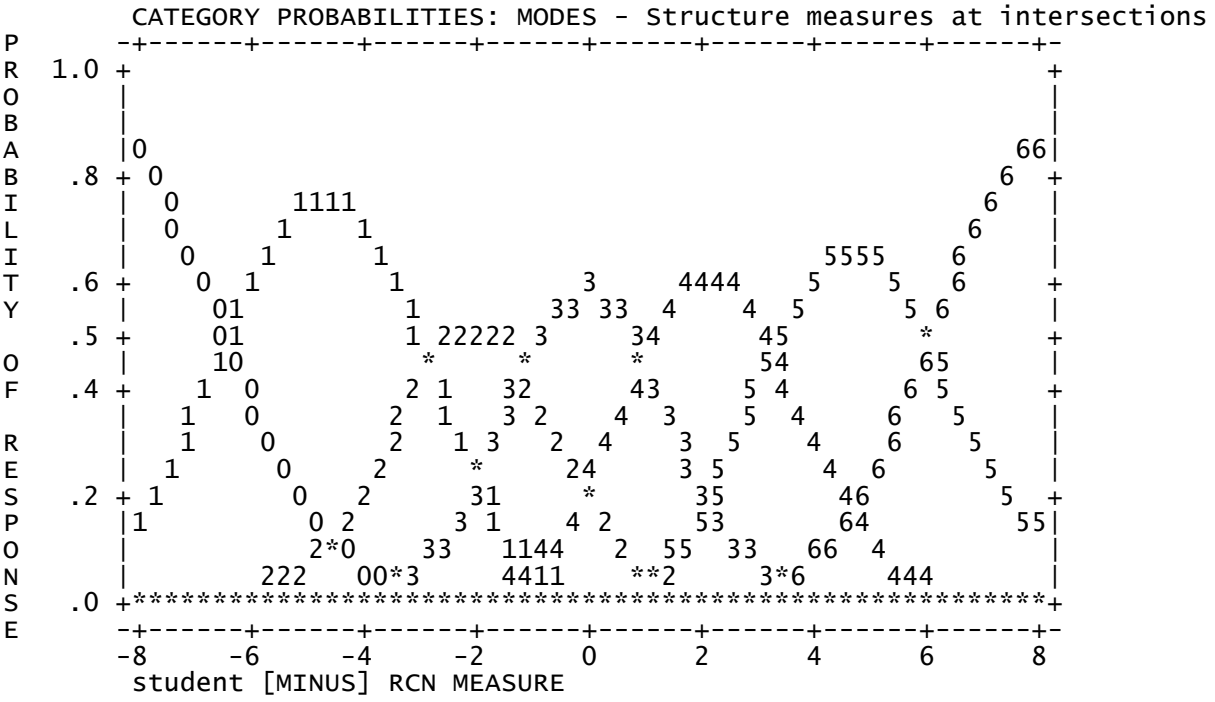


Figure 5. Person-item map for full 23-item, 7-point scale RSAS-RCN.

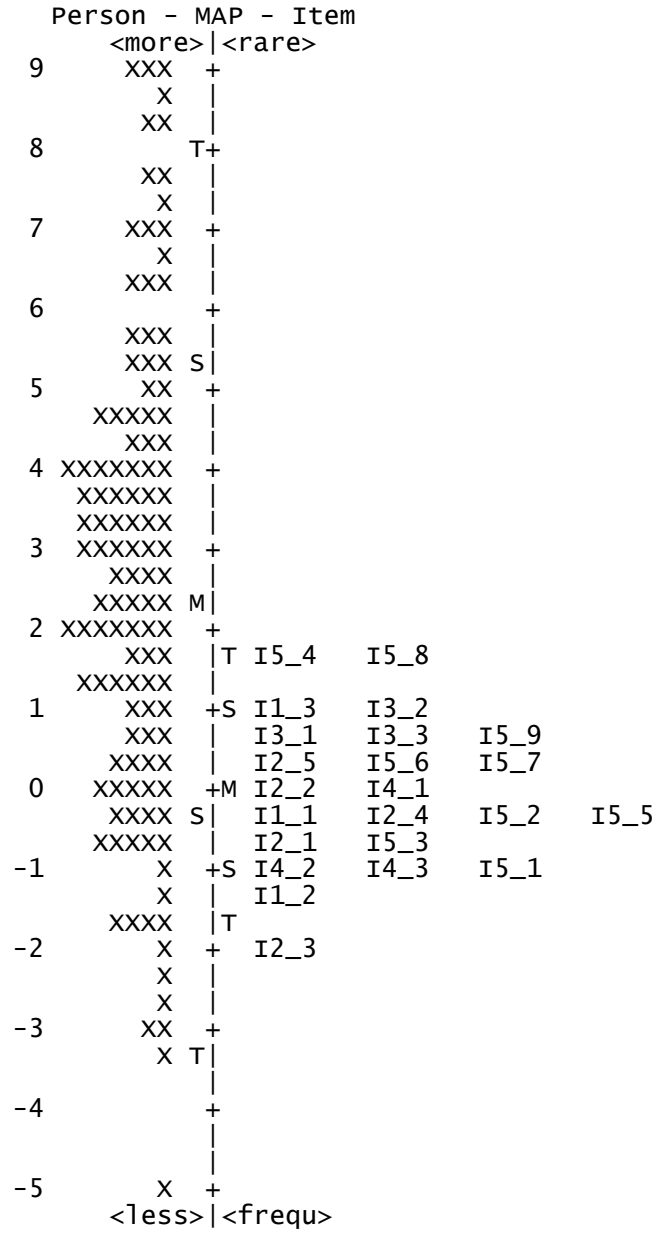


Figure 6. Category probabilities for modified 21-item, 6-point scale RSAS-RCN.

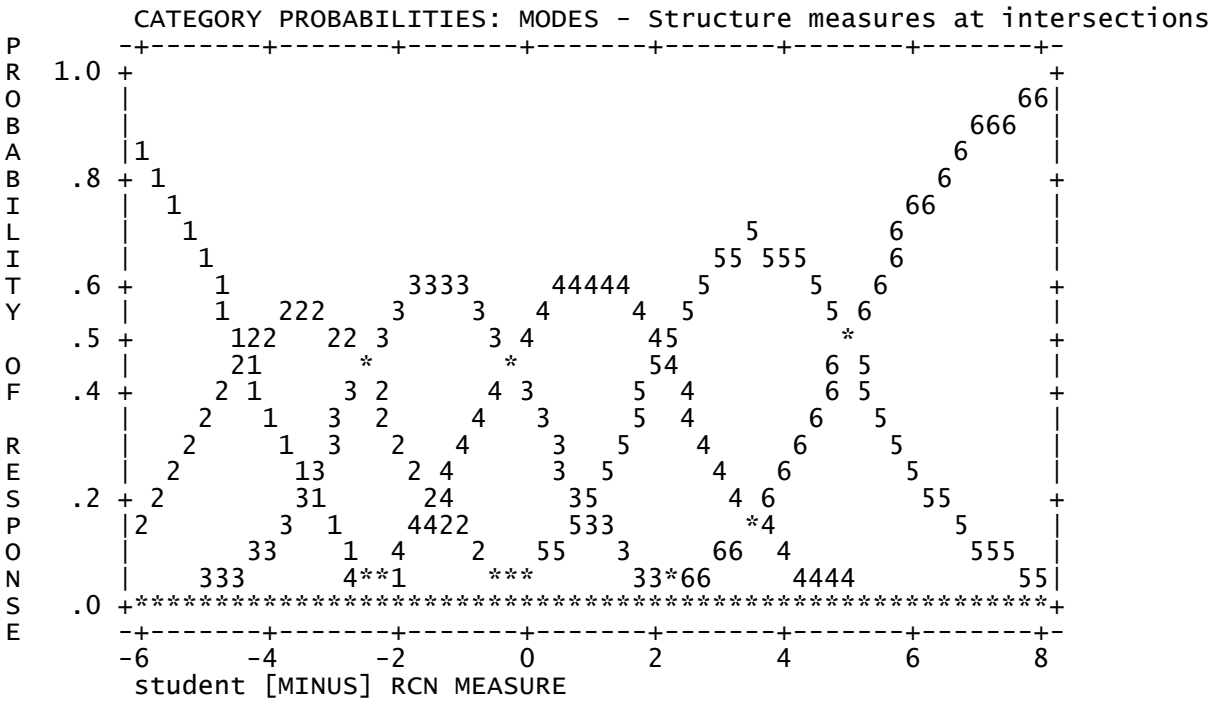


Figure 7. Person-item map for modified 21-item, 6-point scale RSAS-RCN.

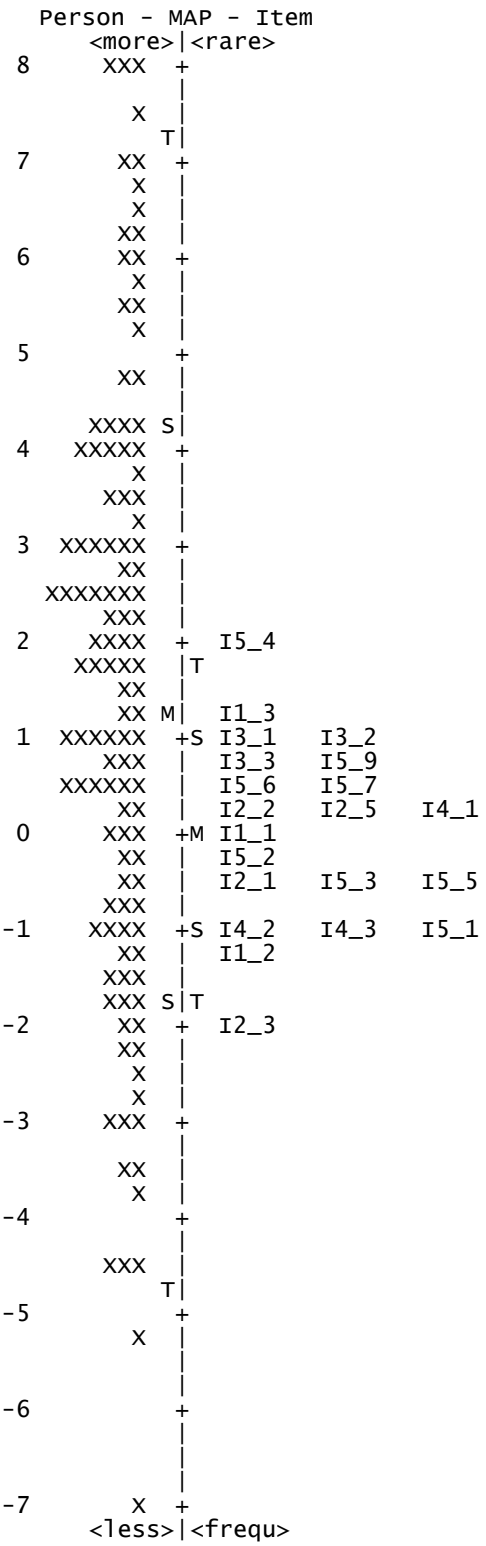
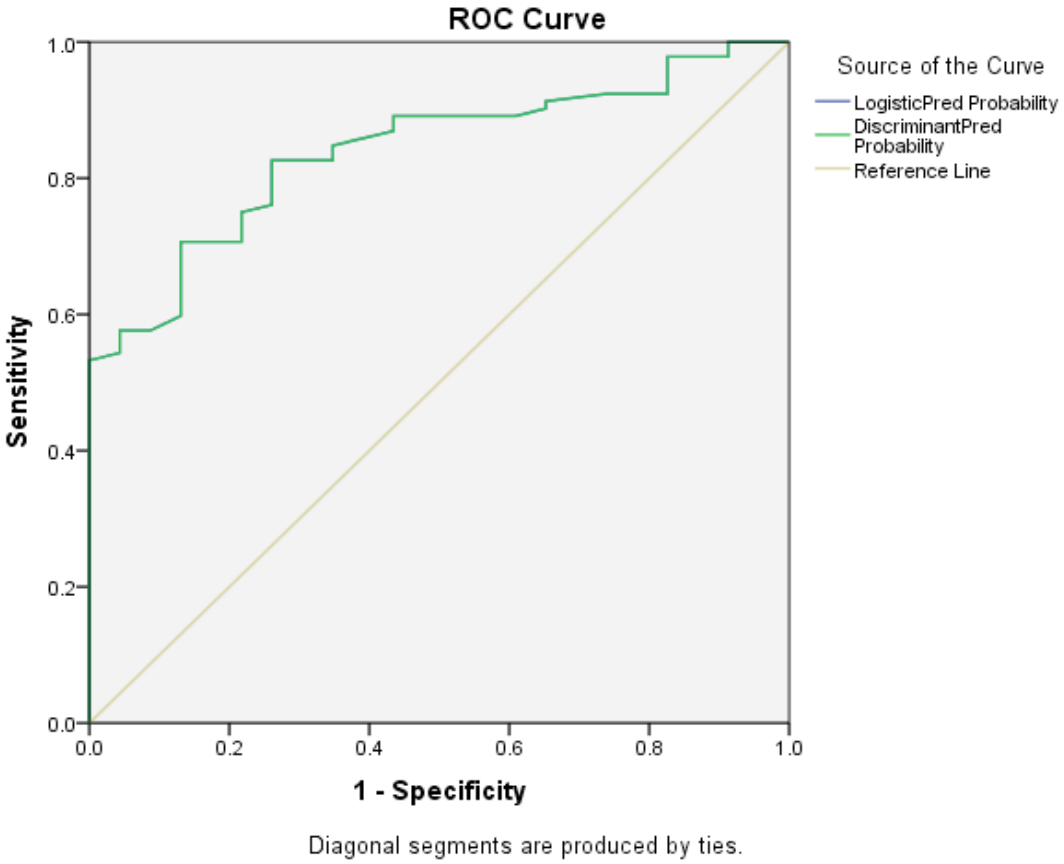


Figure 8. ROC Curve Analysis plot for RSAS-RCN total score to PSSA-Reading "proficient" or "not proficient" scores.



Sarah N. Gebhardt

Lehigh University

6956 N. Wolcott Ave. #3 (616) 648-2750
Chicago, IL 60626 sng209@lehigh.edu

EDUCATION, AWARDS & HONORS

2009 – Present **Ph.D., School Psychology**
Subspecialization in Pediatric School Psychology
Lehigh University, Bethlehem, Pennsylvania
APA accredited and NASP approved program

2007-2009 **M.S. + 30 hours, School Psychology**
Miami University, Oxford, Ohio
NASP approved Program
Thesis Title: Title: Effects of Peer-Monitored Social Skills Training on Indicators of Psychopathology
Thesis Advisor: Dr. T. Steuart Watson
Academic Achievement Award, August 2007

2002-2006 **B.A., Majors: Psychology and Music**
Albion College, Albion, Michigan
Magna Cum Laude with Albion College Honors
The Phi Beta Kappa Society (2006)
Omicron Delta Kappa – Honor Society in Leadership (2005)
Psi Chi – The Honor Society in Psychology (2004)
Alpha Lambda Delta – Honor Society for Freshman (2002)

CLINICAL EXPERIENCE

Aug 2012 – Present **Doctoral School Psychologist Intern**
Illinois School Psychology Internship Consortium – North Suburban Special Education District, Early Childhood Placement
University Supervisor: Christy Novak, Ph.D.
Site Supervisors: Melissa Brown, Ph.D., and Ilene Holt-Turner, Ph.D.

- *Conduct comprehensive psychoeducational assessments, including RtI and functional-based assessments, with students in pre-K through 8th grade*
- *Provide instructional and behavioral coaching and consultation in 2 Pre-K, 1 early childhood, and 1 autism support classroom*
- *Deliver district-wide trainings on topics such as executive functioning assessment and reading assessment/intervention*
- *Participate on district- and school-level RtI leadership teams and grade-level problem-solving teams*
- *Deliver social-emotional learning interventions including class-wide social skills training, video self-modeling curricula, and group and individual counseling*

Aug 2010 – July 2012 **Psychometrician**
Lehigh Psychological Services, Emmaus, Pennsylvania
Supervisor: Daniel Werner, Psy.D.
- *Conducted psychoeducational assessment batteries for children and adolescents*

Sept 2010 – July 2011 **Doctoral Pediatric School Psychology Practicum Student**
Children's Hospital of Philadelphia – Center for Autism Research
University Supervisor: Christy Novak, Ph.D.
Site Supervisor: Lisa Blaskey, Ph.D.
- *Conducted cognitive, language, and achievement testing and structured behavioral observations of children with Autism Spectrum Disorders (ASD)*
- *Participated in research team diagnostic meetings and parental feedback meetings.*

Sept 2009 – June 2011 **Doctoral School Psychology Practicum Student**
Allentown School District, Allentown, Pennsylvania
University Supervisor: Christie Novak, Ph.D.
Site Supervisors: Tama Tamarkin, Psy.D., Deborah Cybuck, M.Ed., and Ilsa Loetzbeier, Ed.S.
- *Conducted comprehensive multidisciplinary psychoeducational evaluations for K-8th grade school students in a large, urban school district*
- *Provided instructional and behavioral consultation to teachers in general education and autism support classrooms*
- *Led social skills counseling groups for kindergarteners and 6th and 7th grade girls*
- *Designed and administered a brief nutrition and wellness program for early elementary students*

Jan 2010 – July 2010 **Doctoral Pediatric School Psychology Practicum Student**
Lehigh Valley Hospital, Allentown and Bethlehem, Pennsylvania
University Supervisor: Patricia H. Manz, Ph.D.
Site Supervisors: Jarred Patten, M.D. and Robert Miller, M.D.
- *Multidisciplinary team member for out-patient pediatric clinics focused on ADHD, pulmonary disease, and complex care*
- *Conducted brief family behavior therapy and clinic-school consultation to aid children and families across home and school settings*

Jan – May 2009 **Master's Level School Psychology Practicum Student**
Loveland School District, Loveland, Ohio
University Supervisor: Raymond Witte, Ph.D.

Jan – May 2009 **Master's Level Counseling Psychology Practicum Student**
St. Nicholas Academy, Deer Park, Ohio
University Supervisor: Susan Mosely-Howard, Ph.D.

Jul 2006 – Jul 2007 **Youth Treatment Specialist**
Wedgwood Christian Services, Grand Rapids, Michigan
Supervisor: Deanna House, M.S.W.
- *Direct care provider at a residential treatment facility for children and adolescents.*

RESEARCH EXPERIENCE

July 2011 – July 2012 **Research Assistant**
National Center on Response to Intervention
Advisor: Edward S. Shapiro, Ph.D.
- *Served as a technical assistant on nation-wide project working to implement Response to Intervention (RTI)*

Aug 2010 – June 2011 **Research Assistant**
College of Education
Lehigh University, Bethlehem, Pennsylvania
Project: RAMP-UP Reading Comprehension Project
Investigators: Mary Beth Calhoon, Ph.D., & Edward S. Shapiro, Ph.D.
- *Worked on a study that investigated the effects of an innovative, year-long reading comprehension curriculum (RAMP-UP) for struggling 6th grade readers at an urban middle school*
- *Responsibilities include revision of curriculum lessons, teacher training, student evaluation, program fidelity checks, and data collection.*

Aug 2009 – June 2011 **Research Assistant**
Center for Promoting Research to Practice
College of Education
Lehigh University, Bethlehem, Pennsylvania
Project: Direct Academic Rating Scale Project
Investigators: Edward S. Shapiro, Sandra M. Chafouleas, Ph.D., & Chris Riley-Tillman, Ph.D.
- *Worked to develop and pilot direct academic rating tools in the areas of reading comprehension and mathematics.*
- *Responsibilities included item development and content validation.*

Project: CBM vs. Computer Adaptive Testing Project
Investigator: Edward S. Shapiro, Ph.D.
- *Worked with two rural elementary schools to compare the utility of a computerized adaptive assessment program with CBM assessments in math and reading*
- *Responsibilities included collecting, entering, and interpreting data*

Aug 2007 – Jul 2009 **Research Assistant**
Department of Educational Psychology
Miami University, Oxford, Ohio
Supervisor: T. Steuart Watson, Ph.D.
- *Completed literature reviews for publications on topics including behavioral tics, compliance, punishment, and identification of Specific Learning Disorders under a response to intervention framework.*

PUBLICATIONS

Shapiro, E. S., & **Gebhardt, S. N.** (2012). Comparing Computer Adaptive and Curriculum-Based Measurement Methods of Assessment. *School Psychology Review, 41,* 295-205.

Watson, T. S., Watson, T., Cole, J. C., & **Gebhardt, S.** (2009). Evidence based practice and learning disabilities. In P. Sturmey & M. Hersen (Eds.), *Handbook of evidence based practice in clinical psychology*. New York: Wiley.

Watson, T.S., Watson, T., & **Gebhardt, S.** (2008). Tics in children: Information for parents and educators. In A. Canter, L. Paige, & S. Shaw, (Eds.), *Helping children at home and school: Handouts from your school psychologist* (3rd ed). Washington, DC: National Association of School Psychologists.

Watson, T.S., Watson, T., & **Gebhardt, S.** (2008). Compliance at home and in the classroom. In A. Canter, L. Paige, & S. Shaw, *Helping children at home and school: Handouts from your school psychologist* (3rd ed). Washington, DC: National Association of School Psychologists.

Watson, T.S., Watson, T., & **Gebhardt, S**. (2008). Classroom management: Punishment. In E. Anderman & L. Anderman (Eds.), *Psychology of classroom learning: An encyclopedia.* Farmington Hills, MI: Thompson.

CONFERENCE PRESENTATIONS

Shapiro, E.S. & **Gebhardt, S.N.** (2013). *Pilot Validation of an Academic Rating Scale of Reading Comprehension.* Paper presented at the annual conference of the National Association of School Psychologists, Seattle, WA.

Gebhardt, S.N., & Shapiro, E.S. (2013). *Comparison of Growth Patterns Across Mathematics Screening Measures.* Poster presented at the annual conference of the National Association of School Psychologists, Seattle, WA.

Shapiro, E.S., Calhoon, M. & **Gebhardt, S.N.** (2013). *Development and Validation of a Teacher Rating Scale for Assessing Reading Comprehension.* Poster presented at the annual conference of the Pacific Coast Research Conference, San Diego, CA.

Gebhardt, S. N. & Shapiro, E. S. (2011). *Comparing Computer Adaptive and Curriculum-Based Measurement Methods of Assessment.* Paper presented at the annual conference of the National Association of School Psychologists, San Francisco, CA.

Shapiro, E. S., & **Gebhardt, S.** (2011). *Comparing a CBM and Computer Adaptive Method as Benchmarks for Assessing Mathematics in Elementary School Students.* Paper presented at the annual conference of the Pacific Coast Research Conference, San Diego, CA.

Gebhardt, S. N., Shaffer, E., Watson, T. S., & Jones, K. (2010). *Effects of Peer-Monitored Social Skills Training on Indicators of Psychopathology.* Poster presented at the annual conference of the National Association of School Psychologists, Chicago, IL.

Shaffer, E., **Gebhardt, S. N.,** Watson, T. S., & Jones, K. (2010). *Effects of Peer-Monitored Social Skills Training Measures of Social Acceptance.* Poster presented at the annual conference of the National Association of School Psychologists, Chicago, IL.

Dumford, N., Watson, T. S., Nickanowicz, C., Bixler, C., & **Gebhardt, S.** (2008). *The Effects of External Rewards on Intrinsic Motivation.* Poster presented at the annual conference of the National Association of School Psychologists, New Orleans, LA.

Gebhardt, S. N. & Keyes, B. (2006). *The effect of race of administrator on children's racial preferences.* Poster presented at the annual conference of the Association for Psychological Science, New York, NY.

PROFESSIONAL ACTIVITIES

Memberships

2007 – Present	Student affiliate, National Association of School Psychologists
2010 – 2012	Student member, Association of School Psychologists in Pennsylvania
2007 – 2009	Student affiliate, Ohio School Psychology Association
2007 – 2009	Student affiliate, South West Ohio School Psychology Association
2006 – 2007	Student member, Association for Psychological Science

Departmental

May 2010 – May 2011 **NASP Student Leader**
Department of School Psychology, Lehigh University, Bethlehem, PA

Aug 2009 – Aug 2010 **Graduate Student Senate Representative**
Department of School Psychology, Lehigh University, Bethlehem, PA

Aug 2009 – Aug 2010 **School Psychology Club Vice President/Co-Chair**
Department of School Psychology, Lehigh University, Bethlehem, PA

References Available Upon Request