

ABSTRACT

Title of Document: VALIDATION OF TEACHER READING RATINGS WITH DIRECT MEASURES OF READING

Julie Adina Grossman, Master of Arts, 2011

Directed By: Professor Deborah L. Speece, Ph.D.
Department of Counseling, Higher Education,
and Special Education

Recent legislation has mandated that students are proficient in reading. Thus, efficient methods of assessment are essential at the present time. Although direct assessments of reading have been shown to be valid in depicting students' skills, they are not efficient methods. It would be cost and time efficient if there were a valid teacher rating instrument. The present study assessed the concurrent and predictive validity of first and fourth grade teacher ratings on Ratings of Overall Reading and Ratings of Reading Problems when compared with several direct measures of reading. Teachers' ratings on Ratings of Overall Reading produced higher validity coefficients than their ratings of the Ratings of Reading Problems. Given that this measure demonstrated the concurrent and predictive validity of teachers' ratings in both first and fourth grades, it is hoped that it can begin to be incorporated into screenings for identifying students experiencing reading difficulties.

VALIDATION OF TEACHER READING RATINGS WITH DIRECT MEASURES
OF READING

By

Julie A. Grossman

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Arts
2011

Advisory Committee:
Professor Deborah L. Speece, Chair
Professor Rebecca Silverman
Professor William Strein

© Copyright by
Julie Adina Grossman
2011

Dedication

I dedicate this thesis to my late grandmother, Elfriede Sonnenschein. She provided me with love and support, and was always willing to lend an ear, laugh at my jokes or spend time playing cards with me when I was in need of a break from my thesis. Although she was not physically present through the completion of my thesis, she certainly was and continues to be present in spirit. Although gone, she is not forgotten. To her, my loving, ever-supportive, gin-rummy partner, I dedicate my master's thesis.

Acknowledgements

I would like to thank the members of my committee, Drs. Deborah Speece (chair), William Strein, and Rebecca Silverman. I appreciate the time you invested as well as the guidance you provided me with, which helped to make this a wonderful and enriching experience. A special thanks to my master's thesis advisor Deborah Speece for continuing to encourage me and answer my endless questions with good cheer and good advice.

I would also like to thank the members of the Reading Research Team as well all of the participants in the two larger studies. A special thanks to Katryna Andrusik for preparing the data set and helping make it available to me.

Last but far from least, I would like to thank my support system – my family. I could not have made it this far without your ongoing support, encouragement, unwavering love, and continuing sense of humor. When I was in need of a break you entertained me. When I was in need of motivation, you helped create a fire underneath me. You helped me to come into my own while still having a strong foundation of support to depend upon.

Table of Contents

Dedication.....	ii
Acknowledgements.....	iii
Table of Contents.....	iv
Chapter 1: Introduction.....	1
Chapter 2: Review of Literature.....	7
Search Methods.....	7
Research Literature.....	8
Concurrent Validity	9
Ratings of general reading ability.....	9
Ratings of specific reading ability	12
Ratings using different analytical methods.....	15
Summary of research on concurrent validity	17
Predictive Validity	18
Summary of research on predictive validity	21
The Present Study	21
Chapter 3: Method.....	23
Participants	23
Measures	24
First grade concurrent validity	24
Phonological awareness.....	25
Word recognition and decoding.....	26
First grade predictive validity.....	28

Word recognition and decoding.....	28
Reading comprehension.....	28
Fourth grade concurrent validity	29
Word recognition and decoding.....	30
Reading comprehension.....	31
Fourth grade predictive validity.....	32
Teacher Ratings	32
Academic Competence (Reading Item).....	32
Reading rating form.....	33
Procedure	33
Teachers	33
Students.....	34
Analytic Plan	35
Preliminary analyses.....	35
Pearson Product Moment Correlation	35
Fisher's z transformation	36
Analyses.....	36
Concurrent and predictive validity	36
Chapter 4: Results.....	41
Assumptions	41
Research Questions.....	41
Research Question 1	42
Concurrent	44

First grade	44
Fourth grade.....	45
Predictive	45
First grade	45
Fourth grade.....	46
Academic competence (reading item)	46
Research Question 2	47
Concurrent	49
First grade	49
Fourth grade.....	49
Predictive	50
First grade	50
Fourth grade.....	51
Academic competence (reading item)	51
Research Question 3	52
Teacher ratings of overall reading	52
Teacher ratings of specific reading problems.....	54
Academic competence (reading item)	55
Chapter 5: Discussion.....	57
Findings	57
Implications	59
Limitations	62
Implications and Future Research.....	63

Conclusion	65
Appendix A.....	67
Appendix B.....	73
References.....	75

List of Tables

Table 1.	First Grade Assessments Divided by Wave	25
Table 2.	Fourth Grade Assessments Divided by Wave.....	29
Table 3.	Correlations for Question 1: Are teacher ratings of 1 st and 4 th graders’ overall reading performance related to direct measures of reading performance and ratings of academic competence?.....	37
Table 4.	Correlations for Question 2: Are teachers’ ratings of the number of specific problems reading of 1 st and 4 th grade children related to direct measures of reading performance and ratings of academic competence? ...	38
Table 5.	Correlations for Question 3: Is there a difference in the strength of the validity coefficients by grade?.....	39
Table 6.	Correlations for Question 1: Are teacher ratings of 1 st overall reading performance related to direct measures of reading performance and ratings of academic competence?.....	43
Table 7.	Correlations for Question 1: Are teacher ratings of 4 th graders’ overall reading performance related to direct measures of reading performance and ratings of academic competence?.....	43
Table 8.	Correlations for Question 2: Are teachers’ ratings of the number of specific problems reading of 1 st grade children related to direct measures of reading performance and ratings of academic competence?.....	47
Table 9.	Correlations for Question 2: Are teachers’ ratings of the number of specific problems reading of 4 th grade children related to direct measures of reading performance and ratings of academic competence?.....	48

Table 10. Correlations for Question 3: Is there a difference in the strength of the validity coefficients of Teachers Overall Ratings by grade?.....	53
Table 11. Correlations for Question 3: Is there a difference in the strength of the validity coefficients of Teachers' Ratings of Number of Problems by grade?.....	54
Table 12. Teacher Background.....	62

Chapter 1: Introduction

Far too many children in the United States of America never succeed in becoming good readers. For example, according to the 2007 National Assessment of Educational Progress, only 33% of 4th graders attained proficient levels in reading (Lee, Grigg, & Donahue, 2007). The longer a child's reading difficulties go without intervening, the further behind the child falls and the less likely later interventions will be effective (Al Otaiba et al., 2009; Donovan & Cross, 2002). Thus, it is important to identify reading difficulties as early as possible. Additionally, early identification of reading difficulties is important because reading serves as a foundation for all other academic subjects. If children are having difficulty reading, it likely will affect their performance in other subjects as well.

Concerns with improving children's reading skills have been longstanding as indicated by the many research studies and federal funds devoted to the issue. Federal involvement dates back to at least 1975 with the passage of PL-94-142. The most recent federal involvement has been NCLB and IDEA. The "No Child Left Behind Act" (NCLB, 2001) holds educators accountable for children's academic success by having the government award federal funding only to those states that have achieved certain educational benchmarks. This legislation emphasizes the need for early identification of children who are displaying academic difficulties.

The *Individuals with Disabilities Education Improvement Act of 2004* (IDEIA, 2004) is the latest reauthorization of a series of laws designed to provide a free and appropriate education for all students. Starting with PL94-142, all children, including those with disabilities (reading disabilities and others), are entitled to a free and

appropriate education. However, prior to IDEIA (2004) in order to qualify for special education services as learning disabled, children needed to display a discrepancy between their cognitive potential and academic achievement. There were many problems with the discrepancy model including its vagueness, lack of applicability for instruction, reliance on extra personnel and, perhaps most importantly, it required that students fail before being assessed. With the passing of IDEIA, the discrepancy rule is no longer required to be used by states when assessing eligibility for special education services. Instead, states now have the option to choose an alternative method, Response to Intervention. Response to Intervention is a more preventative model that systematically applies assessment and increases the intensity of intervention provided to a student rather than immediately assessing that student for eligibility into special education (Batsche et al., 2006). Thus, the Response to Intervention model increases the likelihood that a greater number of students are identified early by implementing universal screening. Universal screening essentially assesses all students for academic difficulties. Although universal screening typically uses direct measures of child performance, it is feasible that teacher ratings of behavior may also be useful.

Direct assessments have been shown to be accurate in determining children's reading skills, but such tests may involve extensive assessment batteries which are labor-intensive. The following two studies demonstrate the large amount of time required to administer direct assessments to children. O'Connor and Jenkins (1999) found that accurate classifications of children at risk at the end of first grade can be made using a battery that requires between 35-65 minutes per student to assess letter

naming fluency, phoneme segmentation, and sound repetition. The assessment battery used by Compton, Fuchs, Fuchs and Bryant (2006) required a five-week time commitment in order to monitor short-term progress of first-grade students' word identification skills. The screening battery used by Compton et al. (2006) differs from that of O'Conner and Jenkins (1999) in that it includes phonemic awareness, rapid naming, oral language, initial word identification, and 5-week progress monitoring of students' word identification level and slope. The researchers found that all of these measures were needed in order to make classifications that met accepted standards of sensitivity and specificity.

Although these screening procedures for reading difficulties are valid, there may be more efficient methods, such as teachers' ratings. Teachers, by virtue of their job, spend a lot of time interacting with children on academic tasks. It would be cost and time efficient if teachers were valid raters, possibly reducing the amount of time children spend in assessments. Additionally, the production of valid teacher rating instruments would allow for a more efficient means of universal screening to identify students who are experiencing difficulties and in need of additional instruction. Such identification is necessary if children are to attain the achievement benchmarks indicated by federal legislation (NCLB, 2001).

Research on the validity of teachers' ratings of students' performance have had mixed findings, which may be due to the use of different methodologies. Studies that have supported the validity of teachers' ratings have involved mainly teachers' ratings of students' behavioral skills, overall academic performance, and specific academic areas including reading. For example, DuPaul and Rapport (1991) found

that teachers provided valid ratings of students' general academic success, behavioral control, and ability to remain focused. Similarly, Demaray and Elliott (1998) found teachers' ratings of students' academic achievement and academic competence (which includes motivation and other items) were strongly correlated with direct assessments of students' skills in mathematics, reading and spelling. On the other hand, Graney (2008) found teachers' ratings of students' progress in oral reading fluency were not significantly correlated with direct measures of students' progress in oral reading fluency. Graney's (2008) findings may differ from DuPaul and Rapport's (1991) and Demaray and Elliott's (1998) because Graney studied the validity of teachers' ratings on students' abilities over time rather than at one point in time.

Research only recently has begun to examine the validity of teachers' ratings of students' reading performance. Accordingly, only a few measures of teachers' ratings of reading performance have been validated. Measures of teachers' ratings of reading performance currently are in demand because once such measures are created and validated, teachers' ratings can then be used as a method for assessing children either alone or in conjunction with other measures or such ratings can be incorporated into a screening battery for reading.

The present study examined the concurrent and predictive validity of a newly created measure of teachers' ratings of students' reading behavior. The proposed study addressed the following questions:

1. Are teacher ratings of first and fourth graders' overall reading performance concurrently and predictively related to direct measures of reading performance and ratings of academic competence (reading item)?

2. Are teachers' ratings of the number of specific reading problems displayed by first and fourth grade children concurrently and predictively related to direct measures of reading performance and ratings of academic competence (reading item)?
3. Is there a grade-related difference in the strength of the validity coefficients?

I expected to find moderate to strong relations ($r \geq .5$) between first and fourth grade teachers' ratings of overall reading performance and norm-referenced tests measuring phonological awareness, comprehension, word recognition and decoding, and overall reading. This was based on prior research showing that teachers' ratings of overall reading performance are moderately to strongly associated with direct assessments (Hopkins et al., 1985; Kenny & Chekaluk, 1993; Gresham et al., 1987). Additionally, similar relationships were expected between teachers' ratings of overall reading and measures of academic competence (reading item) based on findings by Feinberg and Shapiro (2003) and Hecht and Greenfield (2001).

Second, a negative correlation, moderate to strong, was expected between teachers' rating of the number of reading problems and students' scores (first grade: phonological awareness, comprehension, and word recognition and decoding; fourth grade: comprehension, word recognition and decoding, and overall reading performance). Similar associations were expected between teachers' ratings of the number of reading problems and academic competence (reading item; Feinberg & Shapiro, 2003; Hecht & Greenfield, 2001).

Third, research by Kenny and Chekaluk (1993) has suggested that teachers' provide more valid ratings of older children than younger children. In other words, the validity coefficients are higher for older children than younger children. Thus, I expected the correlations between fourth grade teachers' ratings of overall reading abilities and direct assessments of word recognition and decoding to be stronger than the associations between first grade teachers' ratings of overall reading abilities and the same direct assessments.

Chapter 2: Review of Literature

Although valid reading tests exist, some take extensive time to administer. A more efficient way of examining students' reading performance may be through teachers' ratings. Past research has found teachers provide valid ratings of students' behaviors (DuPaul & Rapport, 1991). Other research has assessed the validity of teachers' ratings of reading performance (Hoge & Coladarci, 1989; Hoge, 1983, Perry & Meisels, 1996) and has found that teachers provide valid ratings. However, the research is fairly limited. Existing research on this topic has focused mainly on concurrent validity of teachers' ratings of students' reading performance *or* predictive validity of teachers' ratings of students' reading performance but has not looked at both within a single study.

Search Methods

Three electronic databases, Psych INFO, ERIC, and Academic Search Premier, were searched. Search terms included 'teachers,' 'ratings,' 'ratings of reading,' 'tests,' 'teacher ratings,' and 'reading tests.' To be included in the present review, articles needed to be peer-reviewed and involve elementary school age children, teachers' ratings of reading and a criterion variable of student reading performance. Articles were excluded from the present review if they described teachers predicting students' performance on a specific test rather than teachers' providing ratings on a separate teacher rating scale. Pertinent studies were entered on the ISI Social Sciences Citation Index to find additional relevant studies. The same inclusion and exclusion criteria were applied to all subsequently found articles. The literature search concluded when no new studies continued to be found.

Research Literature

The following section provides a review of 11 studies organized by type of validity. The first eight articles discuss concurrent validity between teachers' ratings of reading and measures of student reading performance. The remaining three articles address predictive validity of teachers' ratings in which teachers' ratings of reading were used to predict students' later reading abilities. Appendix A provides information about each research article, including the title, the authors, the research questions, the measures used, the results, and additional comments.

It should be noted that there is not an agreed upon strength of a validity coefficient that indicates that a teacher's rating is valid. That is, there appears to be no specific number that constitutes acceptable validity. Some (Anastasi & Urbina, 1997) have asserted that there is no minimum validity coefficient indicative of acceptable validity and therefore guidelines from Cohen (1988) should be followed (weak correlation = .1 to .3; moderate correlation = .3 to .5; strong correlation \geq .5). Others have suggested that only coefficients above $r = .4$ to .5 demonstrate acceptable validity (Wood, Garb, & Nezworski, 2007). Coefficients in the literature reviewed below range from $r = .48$, (Teisl, Mazzocco, & Myers, 2001) to $r = .90$ (Farr & Roelke, 1971). Because there is no consensus concerning what an acceptable validity coefficient is, the correlation coefficients presented in this review are described using the following criterion: weak validity coefficients are those below .5, moderate validity coefficients range from .5 to .6, and strong validity coefficients are .6 and above. Both moderate and strong validity coefficients were interpreted as demonstrations of validity.

Concurrent Validity

Eight studies were identified that examined the concurrent validity of teachers' ratings. Of these studies, three focused on teachers' ratings of children's general reading ability, four assessed teachers' ratings of children's specific reading abilities (e.g., decoding, comprehension), and one examined how concurrent validity of teachers' ratings changes when using different analytical methods.

Ratings of general reading ability. Hopkins, George, and Williams (1985), Kenny and Chekaluk (1993), and Gresham, Reschley, and Carey (1987) all asked teachers to provide one overall rating of children's reading performance. Hopkins et al. (1985) had 42 fourth and fifth grade teachers rate 1,032 students' reading performance. Teachers were directed to rate their students' current achievement in five areas, one of which was reading. The five-point scale ranged from poor achievement to excellent achievement. Thus, the teachers' reading rating was based on teachers' response to a single item.

Two weeks after teachers rated students and ranked their reading performance, students were assessed by school district personnel with *The Comprehensive Test of Basic Skills* (CTBS, Form S, Level 2), a norm-referenced test that evaluates the students skills in social studies, reading, math, language arts, and science. No information was provided on what aspects of reading were assessed.

Researchers investigated the relation between teachers' ratings and students' performance on the CTBS. Teachers' ratings of reading performance were strongly related to the standardized reading tests ($r = .73$) suggesting that teachers' ratings of reading ability demonstrate concurrent validity of students' reading abilities.

Kenny and Chekaluk (1993) compared the concurrent validity of 63 teachers' ratings and test-based assessments of 312 kindergartners, first and second graders in a cross-sectional study. Teachers were asked to complete a researcher-developed questionnaire that asked them to rate students' reading performance, cognitive ability, and attentional/behavioral deficits. The 15-item questionnaire asked teachers to indicate on a 3-point scale the extent to which an item applied to each child. Note that only five of the 15 items were about students' reading abilities. In addition to the questionnaire, teachers were asked to categorize each student as being an advanced, average, or poor reader. Students were individually assessed on several published, norm-referenced tests including the Lindamood Auditory Conceptualization test, the Peabody Picture Vocabulary Test-Revised, the Syllable Counting test, the Word String Memory test, the Recall of Designs test, the Memory for Sentences test, and two subtests of the Woodcock Reading Mastery Tests-Revised. These norm-referenced tests measured a range of skills including phonemic awareness, phonemic detection, knowledge of syllables, word attack and word identification, working memory, and receptive vocabulary.

The authors assessed validity by regressing teacher ratings and direct measures on word identification measures to see which served as a better predictor of students' word identification and word attack skills. Scores on the word identification and word attack subtests were added together to create a composite score, Basic Skills (BS). Validity also was assessed by determining the percentage of correct teacher reading skill categories.

Teachers' ratings were moderately correlated with children's Basic Skills score ($r = .57$). Additionally, the researchers found that teachers' ratings of older students produced higher correlational coefficients than teachers' ratings of younger students. This conclusion was based on comparing the percentage of correct teacher categorizations across the three categories of advanced, average, and at-risk reading performance. It is important to note, however, that this study was cross-sectional, not longitudinal; different teachers provided ratings each year, and therefore conclusions about teachers' ratings over time should be interpreted with caution.

An alternative explanation to teachers' ratings of older children being more valid is that more second graders than kindergartners were poor readers. Thus, the higher validity coefficient could be an artifact of the number of poor readers. Evidence for this latter explanation comes from the fact that the percentage of children categorized by teachers as "at risk" was relatively consistent over the three years indicating that teachers identified the same number of children each year as being "at-risk" readers.

In comparison to the two previously described studies, Gresham, Reschley, and Carey (1987) studied the concurrent validity of teachers' ratings of overall reading performance of students relative to their peers as well as relative to grade level expectations. Two hundred students averaging nine years of age were assessed on their verbal intelligence (Wechsler Intelligence Scale for Children-Revised; WISC-R) and reading recognition and comprehension skills (Peabody Individual Achievement Test; PIAT). Teachers used a norm-referenced rating instrument, the Teacher Rating of Academic Performance (TRAP) to rate their students' reading

performance relative to peers (ranging from lowest 10% to highest 10%) and relative to grade level expectations (ranging from well below grade level to well above grade level). Thus, the teachers' reading rating was based on two items. The researchers did not specify the number of teachers participating in this study.

Consistent with prior studies, teachers' ratings of students' overall reading competence were moderately to strongly related to direct assessments of students' reading abilities. Teachers' ratings of students' reading relative to grade expectations were moderately to strongly correlated with direct measures (verbal intelligence: $r = .58$, reading recognition: $r = .62$, reading comprehension, $r = .66$). Strong correlations were found between teachers' ratings of students' reading relative to peers' reading abilities and direct assessments (verbal intelligence: $r = .61$, reading recognition: $r = .67$, reading comprehension, $r = .64$). These findings suggest that teachers provide valid ratings of students' reading abilities when rating relative to peers and to grade level expectations.

Ratings of specific reading ability. The following four studies show that teachers provide valid concurrent ratings of students' specific reading skills. The first study focused on the relation between teachers' ratings and assessments of students' oral reading fluency. The second study related teachers' ratings and students' word analysis skill, vocabulary, and comprehension abilities. In the third study, the researchers looked at the relation between teachers' ratings and direct assessments of students' reading vocabulary, reading comprehension, and verbal intelligence in order to assess the concurrent validity of teachers' ratings. The fourth study looked at

relations between teachers' ratings of decoding, reading accuracy, reading fluency, and reading comprehension and direct assessments of students' oral reading fluency.

Feinberg and Shapiro (2003) investigated the validity of teachers' judgments of students' reading comprehension, decoding, reading fluency and vocabulary. Participants were 30 third to fifth grade students and their 30 teachers; each teacher rated only one student. Students' oral reading fluency was assessed by recording the number of words they were able to read correctly in a limited amount of time when provided with a curriculum-based reading passage.

The teachers rated students on the reading subsection of the Academic Competence Evaluation Scales (ACES) which asked about students' comprehension, word attack, vocabulary, identifying a main idea, and fluency. However, the data reported in the study were limited to relations between teachers' ratings and students' reading fluency. The teachers' ratings were strongly and positively related to students' oral reading fluency ($r = .62$) providing further evidence that teachers' ratings appear to be a valid indicator of reading abilities.

Farr and Roelke (1971) assessed concurrent validity of teachers' ratings of word analysis skill, vocabulary, and comprehension by comparing these ratings with direct assessments of students' skills and reading specialists' ratings. Forty-two fifth graders were assessed on all three components over a two-week period by administering the McGoullough Word Analysis, the Gates McGinitie Reading Test, Survey D, and the California Reading Test, Elementary. Over this same period of time, nine fifth grade teachers and six reading specialists rated the students' skills. Additionally, teachers provided ratings of word analysis skills, vocabulary, and

comprehension twice to ensure reliability of teachers' ratings. The three skills were rated on separate days to minimize possible carry-over effects. Reading specialists were unavailable to be assessed twice, however. Note, no specific information about the questions on the ratings scales was provided.

The associations between teachers' ratings and standardized assessments of these skills ranged between $r = .48$ and $r = .92$ (vocabulary: $r = .92$; comprehension: $r = .59$; word analysis $r = .48$). Similar relations existed between teachers' ratings and reading specialists' ratings of word analysis skill ($r = .48$), vocabulary ($r = .76$), and comprehension ($r = .90$).

In a third study, Sharpley and Edgar (1986) explored the concurrent validity of teachers' ratings by looking at the relationship between such ratings and direct assessments of students' reading vocabulary, reading comprehension, and verbal intelligence. The researchers collected data from 230 third to fifth graders and their teachers. The researchers did not provide the exact number of teacher participants but did mention that all participants were recruited from three schools.

All children were assessed in the early spring on their word knowledge, reading comprehension, receptive vocabulary, and vocabulary intelligence using two published, norm-referenced tests (Progressive Achievement Tests and Peabody Picture Vocabulary Test-Revised). At the same time as students were being assessed, teachers were asked to rate their students' present abilities in reading comprehension, reading vocabulary, and verbal intelligence on a 5-point, investigator-designed scale (1=outstanding...5=well below average). There was one item per reading components

for a total of three items. Teachers were not given any other information before rating their students.

Teachers' ratings of reading vocabulary, reading comprehension, and verbal intelligence were moderately to strongly related to the direct assessments. The relation between teachers' ratings of reading vocabulary and direct assessments of reading vocabulary was $r = .42$, $r = .44$ ¹, and between teachers' ratings of verbal intelligence and the corresponding direct assessment was $r = .41$, $r = .15$. A moderate relationship existed between teachers' ratings of reading comprehension and direct assessments of reading comprehension ($r = .50$, $r = .56$).

Begeny, Eckert, Montarello, and Storie (2008) assessed the relation between teachers' ratings of specific reading abilities and direct assessments of students' oral reading fluency. Ten first through third grade teachers rated 87 of their students' reading abilities, including decoding, reading accuracy, reading fluency, and reading comprehension on an investigator-created measure. The researchers administered curriculum-based passages to students and determined their oral-reading fluency by calculating the number of words read correctly and incorrectly per minute, in grade level passages. The teachers' ratings of general reading were highly, positively correlated with students' word reading fluency ($r = .76$).

Ratings using different analytical methods. Recent research also has focused on how concurrent validity of teachers' ratings changes when using different analytical methods. The following study asserts that correlations over-estimate the validity of teachers' ratings whereas other methods, such as percentage agreements,

¹ Results are reported by gender, with the correlation for boys first, followed by the correlation for girls.

present a more realistic picture. Although these are possible interpretations of the findings, alternative interpretations exist that provide different explanations about the variation in validity of teachers' ratings. One such explanation is that validity of teachers' ratings change based on what aspect of reading is being rated. Further detail about this explanation will be provided later.

Eckert, Dunn, Coddling, Begeny, and Kleinmann (2006) studied the relation between teachers' ratings of reading instructional level and direct assessments of students' oral reading fluency. Instructional level is defined as the percentage of words a child reads correctly in a passage, with frustrational level being less than 93%, instructional level being between 93 and 97%, and independent or mastery level being over 97% (Fuchs & Deno, 1982).

The researchers had two teachers estimate 33 second graders' instructional level in grade level texts as well as students' instructional levels in below grade level texts. Subsequently, researchers assessed students' oral reading fluency when reading below grade level, grade level, and above grade level texts, using curriculum-based measures. The researchers used guidelines set forth by Fuchs and Deno (1982) to determine students' instructional level. No further detail was provided about the specific criteria used to determine students' instructional levels. Eckert et al. (2006) found moderate to strong relations between teachers' estimates of the three instructional levels and direct assessments' determination of instructional level in below grade level ($r = .59$), grade level ($r = .72$), and above grade level texts ($r = .83$).

Teachers' estimates of students' instructional levels in different level texts appeared less valid when percentage agreement analyses were used instead of correlations. Percentage agreement analyses indicated low agreement between teachers' estimates of students' instructional levels in below grade level texts and direct assessments' estimates of students' instructional levels in below grade level texts (Frustrational level, 33%; Instructional level, 20%; Mastery level, 20%). When looking at passages that were at grade level, teachers were more accurate in their estimates of students' reading at a frustrational level than instructional or independent level.

Although the researchers suggested that correlations may overestimate the validity of teachers' ratings, this conclusion cannot be made without ruling out other factors that changed across the two analyses. One factor is the nature of teachers' ratings. For the correlations, teachers' estimates of instructional level in different texts were used whereas for the percentage agreement analyses, teachers' estimates of which passage would be at a child's specific instructional level were used. Perhaps teachers' estimates are more accurate when rating their students' present reading level than when rating a hypothesized reading level.

Summary of research on concurrent validity. The eight studies reviewed above addressed teachers' ratings of overall reading performance as well as teachers' ratings of specific reading skills, such as oral fluency. The studies demonstrated weak-to-strong relations between teachers' ratings and direct assessments of reading abilities. Of the twenty-four validity coefficients reported above, 6 were weak, 5 were moderate, and 13 were strong. Findings also indicated that teachers' ratings of older

children have more validity than ratings of younger children. Additionally, stronger relations were found between teachers' ratings of overall reading performance and direct assessments than between teachers' ratings of specific reading abilities and direct assessments.

Predictive Validity

Predictive validity of teachers' ratings of reading is shown by assessing the relation between teachers' ratings and future measures of students' reading achievement. The length of time between completion of teacher ratings and direct measures varies across studies, ranging from two months to two years. Three studies were identified that examined the predictive validity of teachers' ratings. The studies are organized by the amount of time elapsed between teachers' ratings and direct assessments, going from the least amount of time to the most.

Cabell, Justice, Zucker, and Kilday (2009) studied the validity of teachers' ratings of preschoolers' emerging literacy skills. Forty-four teachers rated the emerging literacy skills (print knowledge, alphabet knowledge, writing) of 209 children using a norm-referenced scale, the Clinical Evaluation of Language Fundamentals – Preschool (CELF Preschool-2 PLRS). Teachers indicated whether students displayed certain behaviors, such as picking up a book and flipping through the pages, identifying alphabet letters, and knowing familiar words. The children ranged in age from 40 to 68 months and were racially and ethnically diverse. A majority of the children lived in low-income families. About three months after teachers made their ratings, the researchers returned to assess children's print-concept knowledge, alphabet knowledge, and emergent writing.

The teachers' ratings were correlated with children's print knowledge ($r=.48$), alphabet knowledge ($r=.60$), and writing ($r=.48$). These findings suggest that teachers' ratings are weak to strong predictors of academic performance assessed three months later. However, it is important to note that ratings of emergent literacy skills may differ from ratings of reading skills. One reason for this may be that different skills are assessed with emerging readers than with older children who can already read.

A study by Teisl, Mazzocco, and Myers (2001) assessed the predictive validity of kindergarten teachers' ratings of reading performance by determining the relation between these ratings and students' future reading performance. The researchers gathered teachers' ratings of 234 kindergartners ranging from five to seven years of age, with the average being 5.9 years. No further detail was provided concerning the inclusion of seven-year old kindergarteners.

At the end of the school year, teachers were provided with a norm-referenced rating instrument and asked to rate their students' reading performance on a 5-point scale, ranging from below average to above average. A year later, the researchers assessed the same students' decoding skills using the Letter-Word Identification subtest of the Woodcock Johnson-Revised (WJ-R). There was a positive relationship ($r =.48$) between teachers' ratings of kindergartner's reading performance and direct assessment of reading. The findings by Teisl et al. (2001), suggest that teachers' rating of the reading skills of students', even as young as kindergarteners, are valid predictors of reading performance a year later.

Hecht and Greenfield (2001) assessed both the concurrent and predictive validity of teachers' ratings of reading performance. The students in this study were first and third graders. In the spring of 1st grade, 21 teachers rated 170 students using the Academic Competence subscale of the Social Skills Rating System. At the same time, these first graders also were assessed on their skills with decoding, word recognition, reading comprehension, phonemes, print knowledge, and receptive language with the Letter-Word Identification and Passage Comprehension subtest of the Woodcock Johnson Tests of Achievement – Revised (Form B), the Yopp-Singer Segmentation Test, the Roswell-Chall Auditory Blending Test, the Stones-Concept About Print Test, and the Peabody Picture Vocabulary Test – Revised (Form M). The children were assessed again on their decoding and comprehension skills two years later when they were in third grade.

The researchers found a range of positive correlations between teachers' ratings reading performance and direct assessments: phoneme segmentation ($r = .47$), phoneme blending ($r = .52$), print knowledge ($r = .60$), receptive vocabulary ($r = .51$), word recognition ($r = .75$), and reading comprehension ($r = .68$). In third grade, strong correlations were found between teachers' ratings of reading performance and word recognition ($r = .71$) and reading comprehension ($r = .70$). These results suggest that teachers' ratings of reading performance display both concurrent and predictive validity.

The researchers mention a key limitation of their study was that the sample of students was restricted to those classified as living in poverty. It is important to try and replicate these findings with a more heterogeneous sample to see if there are

achievement differences or other indicators that make teachers more valid raters with that population.

Summary of research on predictive validity. Three studies were reviewed and all provided support for the predictive validity of teachers' ratings of oral fluency, decoding, comprehension, print knowledge, and word recognition. Teachers' ratings were valid predictors of future reading performance assessed between three months and two years later as shown by the moderate-to-strong correlations found between teachers' ratings of reading and later direct assessments. Of the twelve correlations presented above, four were weak, two were moderate, and six were strong. The reviewed studies included preschool through third grade students. The findings suggest that the correlations of teachers' ratings of emerging reading and direct assessments are weaker than correlations between teachers' ratings of older students and direct assessments.

The Present Study

The present study extended current research on the validity of teachers' ratings in four ways. First, the present study included teachers' ratings of both overall reading performance and specific reading abilities. Prior studies have included one or the other, but not both. By having teachers rate overall reading performance and identify specific problem areas for below grade level readers, the present study provided more insight into the extent of problems that below grade level readers experience. Similarly, the present study assessed both concurrent and predictive validity. Prior studies have included one or the other, but often not both. Third, the present study focused on a greater range of reading skills than assessed in previous

studies. Previous studies have determined validity of teachers' ratings by determining the associations between teachers' ratings and future direct assessments of oral fluency, decoding, comprehension, print knowledge, and word recognition. The current study extended prior research by examining these same constructs within a single sample. Fourth, few studies have addressed whether the validity of teachers' ratings differs for younger and older students. The present study compared ratings for younger and older students as well as included older students than have been previously researched. The data from this study came from two larger, ongoing studies of children's reading abilities conducted by Dr. Deborah Speece.

Chapter 3: Method

Participants

The data in the current study came from two larger studies (PI: Dr. Deborah Speece) which consider the identification of children at risk for reading problems and the effectiveness of an intervention for elementary school students at risk for reading difficulties. The data in the present study were from the fall and spring of the first year of the larger longitudinal studies.

Participants in the present study were first and fourth graders and their teachers. Three hundred sixty-seven first grade students from 11 parochial schools and 16 classrooms were invited to participate. Two hundred fifty-seven (70%) parents of first graders gave permission; all these children were included in the study sample. Only children who had complete data were included in the present analysis resulting in a sample of 243 children in the current study. The present sample included 113 female (47%) and 130 male (53%) children with a mean age of 6.56 years ($SD = .32$) prior to the beginning of data collection. Based on parents' reports of students' race and/or ethnicity, about 80% of the sample was Caucasian, 7% African American, 7% Asian, 3% Hispanic, less than 1% American Indian, and approximately 2% reported more than one race. The majority of the students (96%) spoke English as their first language. Twenty-one percent of mothers did not report on their educational level. Of those who did report on their educational level, fewer than 1% of mothers had no high school degree, 42% had a high school degree, 45% had a college degree, and 12% had graduate or professional training.

Three hundred ninety-eight fourth graders from 15 parochial schools in the same school system and 20 classrooms were invited to participate. Two hundred thirty five (59%) parents provided consent. After accounting for a student transferring schools, a student sending in the permission form too late, and refusal of students to participate, the final sample included 230 fourth graders. As of late fall, the mean age of the participants was 9.45 years ($SD = 0.33$ years). Seventy-two percent of the students were Caucasian, 18% were African-American, 1% were Asian, and 4% were bi-racial. No information about race/ethnicity was provided for about 3% of the students. All students spoke English as their primary language. About 15% of the mothers had a high school diploma, 33.7% had some college education, 27.6% had a college degree, and 21.6% had a professional or graduate degree.

Additional information was collected from the 16 first grade teachers and 17 fourth grade teachers in the study. About 94% of the first grade teachers were female and 87.5% held teacher certification. They had been teaching a mean of 19.8 years ($SD = 13.2$). All of the fourth grade teachers were female and 52.9% were certified. They had been teaching a mean of 11.4 years ($SD = 10.4$).

Measures

The assessments were administered in several waves across the school year with some of the waves including individual assessments and others including group assessments

First grade concurrent validity. Measures of phonological awareness, decoding and word recognition were used to establish concurrent validity of teachers' ratings (see Table 1).

Table 1

First Grade Assessments Divided by Wave

First Grade			
Wave	Test Name	Type of Validity	
		Concurrent	Predictive
1	WIF	**	
2	WRMT	**	
2	CTOPP	**	
2	TOWRE	**	
4	WRMT		**
4	CRAB		**
4	PRF		**

Phonological awareness.

Comprehensive Test of Phonological Processing (CTOPP; Wagner et al., 1999). The CTOPP assesses children’s phonological awareness, phonological memory, and rapid naming abilities. One of the core subtests for the Phonological Awareness composite is Elision which assesses one’s ability to segment spoken words into smaller parts. First graders were given the entire 20-item subtest or as much as they could complete before reaching the ceiling rule of three consecutive, incorrect responses. They were assessed on their abilities to say a word and then to say the same word once certain sounds had been eliminated. The raw score is the number of correctly pronounced words after the phonemes have been deleted.

Wagner et al. (1999) report the internal consistency for the Elision subtest was .92 for six-year olds and .89 for nine-year olds. The test-retest reliability for the

Elision subtest for five- to seven-year olds was .88 ($n=32$) and .79 for eight- to seventeen-year olds ($n=30$). The inter-rater reliability was .96 for five- to six-year olds and .99 for seven-year olds and older. The publisher reported concurrent and predictive partial correlations after controlling for age. Concurrent and predictive validity were determined by correlating the Elision subtest with Word Attack and Word Identification subtests from the Woodcock Reading Mastery Tests-Revised ($N = 73$). Concurrent validity was found to be .74 and .53, respectively and predictive validity was found to be .72 and .68, respectively.

Word recognition and decoding.

Test of Word Reading Efficiency (TOWRE; Torgeson, Wagner, & Rashotte, 1999). The TOWRE assesses students' abilities to pronounce words accurately and fluently using two subtests, Sight Word Efficiency and Phonemic Decoding Efficiency. For the Sight Word Efficiency, students were given a list of words that started out easy and got increasingly hard. Students were told to read as many words as fast as they could for forty-five seconds. This procedure then was repeated with a second list of words. The raw score was the total number of words correctly read within the forty-five seconds. The Phonemic Decoding Efficiency was the same as the Sight Word Efficiency subtest with the only difference that the lists were comprised of nonwords.

Torgeson et al. (1999) report the alternative form coefficient for Forms A and B for ages six through nine with a range of .93 to .97 for both Sight Word Efficiency and Phonemic Decoding Efficiency. This statistic was used in place of internal consistency because it was a speeded test. The test-retest reliability was reported for

ages 6 through 9 for both Sight Word Efficiency (Form A: .97; Form B: .96) and Phonological Decoding Efficiency (Form A: .90; Form B: .90). The inter-rater reliability was based on data for students between first grade and twelfth grade and was .99 for both subtests. Concurrent validity was reported for first and fourth graders by correlating the TOWRE with the Woodcock Reading Mastery Tests-Revised. For first graders, the concurrent validity was .89 when compared with the Word Attack subtest of the WRMT-R and .92 when compared with the Word Identification subtest of the WRMT-R. For fourth graders, the concurrent validity as compared with the Word Attack subtest was .87 while the concurrent validity when compared with the Word Identification subtest was .89.

Woodcock Reading Mastery Test (WRMT; Woodcock, 1998). The Woodcock Reading Mastery Test is comprised of two subtests, Word Attack and Word Identification. The former evaluates students' accuracy in decoding pronounceable nonsense words; the latter evaluates students' word recognition. Both subtests require students to decode/recognize individually presented words (nonsense words) that increase in difficulty as the task progresses. Raw scores for each subtest are based on the number of words read correctly. Woodcock (1998) used well-known reading measures to establish concurrent validity of .98 for both subtests.

Word Identification Fluency (WIF). First graders' abilities to quickly identify words were assessed using Word Identification Fluency (WIF) measure (D. Compton, personal communication, 2003). This measure is comprised of two lists of grade-level words, each containing 50 high frequency words. Students were given 1 minute to read as many words as they could. Responses were considered incorrect if the words

were mispronounced, skipped, or uttered after 3 seconds had passed. The number of correct responses in each passage were summed together to create an overall score. The alternate test-form/stability coefficient after a 2-week interval was .88 (Fuchs, Fuchs, & Compton, 2004).

First grade predictive validity. Measures of word recognition and reading comprehension were used to establish predictive validity of teachers' ratings (see Table 1).

Word recognition and decoding.

Passage Reading Fluency (PRF; Fuchs, Hamlett, & Fuchs, 1990). To assess oral reading fluency, 1st graders were individually administered a curriculum-based measure, Passage Reading Fluency (PRF; Fuchs, Hamlett, & Fuchs, 1990). Students were provided with two grade-level narrative passages and given a minute to read as much of each passage as they could. The mean number of words read correctly per minute was calculated. Both test-retest and alternate forms reliability are high ($r > .90$) across studies, and criterion validity is strong (Deno, 1985; Fuchs & Fuchs, 1992; Marston, 1989).

Woodcock Reading Mastery Test (WRMT; Woodcock, 1998). This measure was described previously in the section on concurrent validity

Reading comprehension.

Comprehensive Reading Assessment Battery - Comprehension (CRAB; Fuchs, Fuchs, & Hamlett, 1988). The Comprehensive Reading Assessment Battery-Comprehension assesses students' fluency and comprehension abilities. Students were presented with two 400-word stories written at the 1.5 grade level. Students

were given three minutes to read each story and then asked to respond to ten questions about the story. The student’s fluency (average number of words read correctly per minute) and comprehension (questions answered correctly) during two trials were used as student’s scores. The fluency and comprehension scores show excellent test-retest reliability ($r = .91$) and concurrent criterion related validity of .91 to .92 with the Stanford Achievement Test (SAT) (Fuchs, Fuchs, & Maxwell, 1988).

Fourth grade concurrent validity. Measures of word recognition and reading comprehension were used to establish concurrent validity of teachers’ ratings (see Table 2).

Table 2

Fourth Grade Assessments Divided by Wave

Fourth Grade			
Wave	Test Name	Type of Validity	
		Concurrent	Predictive
1	PRF	**	
1	TOWRE	**	
1	WIF	**	
1	WJ III	**	
2	GATES	**	
2	MAZE	**	
5	GATES		**
5	MAZE		**
6	WJ III		**
6	WIF		**
6	PRF		**

Word recognition and decoding.

Test of Word Reading Efficiency (TOWRE; Torgeson, Wagner, & Rashotte,, 1999). The TOWRE was previously described in the first grade section on concurrent validity.

Woodcock Johnson Tests of Achievement, Third Edition (WJ III; Woodcock, McGrew, Mather, & Schrank, 2001). Two subtests of the Woodcock Johnson Tests of Achievement, Third Edition (Woodcock, McGrew, Mather, & Schrank, 2001) were administered, Word Attack and Word Identification. The former evaluates students' accuracy in decoding pronounceable nonsense words; the latter evaluates students' word recognition. Both subtests require students to decode/recognize individually presented words (nonsense words) that increase in difficulty as the task progresses. Raw scores for each subtest are based on the number of words read correctly. The split-half reliability coefficients for nine-year-old children assessed with the Word Identification subtest and the Word Attack subtest are .94 and .89, respectively.

Passage Reading Fluency (PRF; Fuchs, Hamlett, & Fuchs, 1990). The same measure of passage reading fluency was used with fourth graders as was used with first graders. Again, both test-retest and alternate forms reliability are high ($r > .90$) across studies, and criterion validity is strong (Deno, 1985; Fuchs & Fuchs, 1992; Marston, 1989).

Word Identification Fluency (WIF; Speece et al., in press). WIF is an individually administered curriculum-based measure of word reading fluency developed by Speece et al. (2010). The development of WIF was based on a procedure developed by Compton (described in Speece et al., 2010). The words on

WIF were randomly selected from the Educator's Word Frequency Guide (Zeno, Ivens, Millard, & Duvvuri, 1995). The Educator's Word Frequency Guide is based on a large word frequency study, and provides information on the frequency of words and in what grade students are likely to encounter specific words. Parallel probes, each with 80 words representing a range of frequency levels, were created. The variable of interest was the mean number of words students read correctly in one minute over two trials. Speece et al. (2010) found the parallel-forms reliability coefficient was .92 with a sample of fourth graders. Validity coefficients with the WJ-III Word Identification subtest ($r = .68$), TOWRE Sight Word Efficiency ($r = .86$), and PRF ($r = .78$) are strong.

Reading comprehension.

Gates-MacGinitie Reading Test, Fourth Edition (GMRT). Fourth grade students were assessed on their comprehension abilities through a group administration of the reading comprehension subtest of the GMRT (MacGinitie, MacGinitie, Maria, & Dreyer, 2000). Thirty-five minutes were allotted for students to read narrative and expository passages and answer multiple choice questions about each passage. Form S was administered in the fall and Form T was administered in the spring. MacGinitie et al. (2000) report strong internal consistency, with both the alternate form reliability and the internal consistency being above .90 for fourth grade students.

Maze (Fuchs, n.d.; Fuchs & Fuchs, 1992). The Maze was group-administered to fourth graders (Fuchs, n.d.; Fuchs & Fuchs, 1992) to assess students' abilities in comprehending sentences and correctly choosing words that belong in a sentence.

The students were presented with a reading passage, of which only the first sentence was intact. After the first sentence, every seventh word was deleted and replaced with three choices. Students were given two minutes to complete as many choices as possible. Students completed two reading passages. The mean of the number of correct choices was converted to items correct per minute. Test-retest reliability for the Maze was over .90 with second graders (Guthrie, Siefert, Burnham, & Caplan, 1974).

Fourth grade predictive validity. The Word Identification and Word Attack subtests of the Woodcock Johnson Tests of Achievement, Third Edition (WJ III), Passage Reading Fluency (PRF), and the Word Identification Fluency (WIF) measures were used to assess students' word recognition and decoding skills; the Gates-MacGinitie Reading Test, Fourth Edition and the Maze were used to assess reading comprehension. These five tests were all described above in the fourth grade section labeled concurrent validity. See Table 2 for the measures used to assess the predictive validity of fourth grade teachers.

Teacher Ratings

Academic Competence (Reading Item). The Social Skills Rating System (SSRS; Greshman & Elliot, 1990) is a multi-rater assessment pertaining to student social behaviors and academic competence that can affect academic performance. First and fourth grade teachers were asked to complete the 9-item Academic Competence subtest of the SSRS. The present study only included the one item about reading competence from the SSRS. Teachers were asked to rate a student's reading abilities relative to his or her classmates on a five-point scale (1 = the lowest 10%; 2

= the next lowest 20%; 3 – the middle 40%; 4 = the next highest 20%; 5 = the highest 10%). The items being rated by teachers pertained to their perceptions of students' reading achievement, cognitive skills, and academic motivation. Gresham and Elliot (1990) report an internal consistency of .95. The test-retest reliability reported was .93 with four weeks between administrations of the test. Gresham and Elliot (1990) suggest that the SSRS demonstrates criterion validity as shown when correlating scores from the SSRS with the Harter Teacher Rating Form ($r = .63$; $N = 243$).

Reading rating form. This form was developed by the researchers (Speece et al., 2010). Teachers rated students' overall reading ability on a five-point rating scale (Rating of Overall Reading; RROR). A score of one or two indicated a student was below grade level, a score of three to five indicated a student was on or above grade level. Teachers who provided a rating of a one or two were asked to indicate specific areas of difficulty for a student. The possible choices were: decoding, vocabulary, fluency, comprehension, and motivation. The number of problems indicated by the teacher were then summed together to create a Rating of Reading Problems (RRPR). Based on a sample of fourth graders, analyses reported validity coefficients for the teachers' Rating of Overall Reading and GMRT and Maze ($r = .68$ and $r = .63$, respectively). Initial analyses reported validity coefficients for the teachers' rating of the number of problems and the GMRT and Maze of $-.50$ and $-.49$, respectively (Speece et al., 2010).

Procedure

Teachers. In November of 2006, first and fourth grade teachers received a packet of forms including a Teacher Background form, the Reading Rating Form and

the Academic Competence (reading item) for each student who had returned a parent consent to be in the study. The teachers had one week to complete these forms before a graduate research assistant came to pick them up. The only instructions that teachers were given regarding the Reading Rating Form was to rate every student's overall reading skills. If the teacher had given an overall rating of a one or two (below grade level), the teacher was instructed to check all of the relevant problem areas. Otherwise, they were informed to leave the number of problem areas blank. No further instructions were provided.

Students. Student data were collected by graduate students who were trained to a 90% accuracy criterion (administration and scoring) on all measures before testing began. Also, on-site fidelity checks were made throughout the year. The assessments were administered in several waves across the school year with some of the waves including individual assessments and others including group assessments. Children were tested in the same order within waves to maintain equal spacing of assessments. In general, Wave 1 was collected between November and mid-December, Wave 2 from December through January, Wave 3 in late January through February, Wave 4 from April to mid-May, and Wave 5 was collected in May. Waves 1, 3, and 6 were administered individually while waves 2, 4, and 5 were administered in a group setting.

For the present study, first grade data collected in waves 2 and 4 were used and fourth grade data collected in waves 1, 2, 5, and 6 were used. Table 1 provides more information about which measures were administered during which waves and which measures were used to assess concurrent validity versus predictive validity.

Analytic Plan

Preliminary analyses. Before performing the analyses, assumptions underlying the use of the Pearson Product Moment Correlation and the Fisher's z transformation were examined.

Pearson Product Moment Correlation. The most important assumptions of the Pearson Product Moment Correlation are linearity, independence, and the measurement scale. Violations of these assumptions can result in an underestimation of the relation between the two variables being measured (Hinkle, Wiersma, & Jurs, 2003). Thus, violations can result in a more conservative assessment. Data were examined for linearity. Additionally, the Pearson Product Moment Correlation was only used with data measured on an interval or ratio scale. All of the measures with the exception of one are measured on an interval scale. The teachers' ratings of the students' overall reading ability were measured on an ordinal scale and thus were computed using a Spearman Rho correlation.

To check the linearity of these relationships, I plotted the variable to examine the visual relationship between each of the teachers' ratings of reading (overall and number of problems) with each of the direct assessments. Additionally, I also created a scatter plot between each of the teachers' ratings of reading and their ratings of academic competence (reading item). All scatter plots were examined to make sure a linear relationship existed between all direct assessments and teachers' ratings. Linear relationships were assumed if the data points formed an elliptical, rather than a circular, pattern.

Given the large number of correlations produced in the present study, the possibility of family wise error is acknowledged. However, given that the present study was a validity study in which the emphasis was on the absolute value of the coefficient rather than the statistical significance, the Bonferroni adjustment was not considered.

Fisher's z transformation (Fisher, 1970). The assumptions present in the Fisher's z transformation are the same as those included in the Pearson Product Moment Correlation analyses. Therefore, by meeting the above assumptions, I had already met all assumptions for the Fisher's z transformation. The reason for this is that before transforming the statistic into a z statistic, it will be an r statistic and thus all the assumptions for the Pearson Product Moment Correlation had to have been met.

Analyses.

Concurrent and predictive validity. Concurrent validity was based on students' fall test scores while predictive validity was based on students' spring test scores.

1. Question 1: Are teacher ratings of first and fourth graders' overall reading performance concurrently and predictively related to direct measures of reading performance and ratings of academic competence (reading item)? Spearman *rho* correlations were calculated between teachers' ratings on Rating of Overall Reading and all direct assessments. See Table 3 for details on the specific direct assessments used in these analyses. In addition, the correlation was calculated

between teachers' ratings on Ratings of Overall Reading and teachers' rating of academic competence (reading item).

Table 3

Correlations for Question 1: Are teacher ratings of 1st and 4th graders' overall reading performance related to direct measures of reading performance and ratings of academic competence?

1 st Grade	
Concurrent Validity	Predictive Validity
Teacher Overall and Student CTOPP	Teacher Overall and Student PRF
Teacher Overall and Student TOWRE	Teacher Overall and Student WRMT
Teacher Overall and Student WRMT	Teacher Overall and Student CRAB
Teacher Overall and Student WIF	
Teacher Overall and Teacher SSRS	
4 th Grade	
Concurrent Validity	Predictive Validity
Teacher Overall and Student TOWRE	Teacher Overall and Student WJ III
Teacher Overall and Student WJ III	Teacher Overall and Student PRF
Teacher Overall and Student PRF	Teacher Overall and Student WIF
Teacher Overall and Student WIF	Teacher Overall and Student GMRT
Teacher Overall and Student GMRT	Teacher Overall and Student MAZE
Teacher Overall and Student MAZE	
Teacher Overall and Teacher SSRS	

- Question 2: Are teachers' ratings of the number of specific reading problems displayed by first and fourth grade children concurrently and predictively related to direct measures of reading performance and ratings of academic competence (reading item)?

Pearson Product Moment Correlations were calculated between teachers' ratings of the number of problems and all direct assessments. See Table 4 for details on specific direct assessments used in these analyses. In addition, the correlation was calculated between teachers' ratings of the number of problems and teachers' ratings of academic competence (reading item).

Table 4

Correlations for Question 2: Are teachers' ratings of the number of specific problems reading of 1st and 4th grade children related to direct measures of reading performance and ratings of academic competence?

1 st Grade	
Concurrent Validity	Predictive Validity
Teacher Number of Problems and Student CTOPP	Teacher Number of Problems and Student PRF
Teacher Number of Problems and Student TOWRE	Teacher Number of Problems and Student WRMT
Teacher Number of Problems and Student WRMT	Teacher Number of Problems and Student CRAB
Teacher Number of Problems and Student WIF	
	Teacher Number of Problems and SSRS
4 th Grade	
Concurrent Validity	Predictive Validity
Teacher Number of Problems and Student TOWRE	Teacher Number of Problems and Student WJ III
Teacher Number of Problems and Student WJ III	Teacher Number of Problems and Student PRF
Teacher Number of Problems and Student PRF	Teacher Number of Problems and Student WIF
Teacher Number of Problems and Student WIF	Teacher Number of Problems and Student GMRT

3. Question 3: Is there a grade-related difference in the strength of the validity coefficients?

The Fisher's z statistic was applied to the data to assess whether teachers provide more valid ratings of older students than of younger students (Fisher, 1970). By applying this statistic, I was able to assess whether two correlations were significantly different from one another. Only correlations with $p < .05$ were considered statistically significant. The Fisher's z statistic was applied only to correlations that included direct assessments given to both first and fourth graders. Table 5 indicates the correlations that were compared using the Fisher's z statistic.

Table 5

Correlations for Question 3: Is there a difference in the strength of the validity coefficients by grade?

1 st Grade and 4 th Grade Concurrent Validity	
Teacher Overall Ratings	Teacher Number of Problem Ratings
1 st Grade Teacher Overall and Student TOWRE and 4 th Grade Teacher Overall and Student TOWRE	1 st Grade Teacher Number of Problems and Student TOWRE and 4 th Grade Teacher Number of Problems and Student TOWRE
1 st Grade Teacher Overall and Student WIF and 4 th Grade Teacher Overall and Student WIF	1 st Grade Teacher Number of Problems and Student WIF and 4 th Grade Teacher Number of Problems and Student WIF
1 st Grade Teacher Overall and Student WRMT	1 st Grade Teacher Number of Problems and

and 4 th Grade Teacher Overall and Student WJ III	Student WRMT and 4 th Grade Teacher Number of Problems and Student WJ III
1 st Grade and 4 th Grade Predictive Validity	
Teacher Overall Ratings	Teacher Number of Problem Ratings
1 st Grade Teacher Overall and Student PRF and 4 th Grade Teacher Overall and Student PRF	1 st Grade Teacher Number of Problems and Student PRF and 4 th Grade Teacher Number of Problems and Student PRF
1 st Grade Teacher Overall and Student WRMT and 4 th Grade Teacher Overall and Student WJ III	1 st Grade Teacher Number of Problems and Student WRMT and 4 th Grade Teacher Overall and Student WJ III
1 st and 4 th Grade Academic Competence	
Teacher Overall Ratings	Teacher Number of Problem Ratings
1 st Grade Teacher Overall and SSRS and 4 th Grade Teacher Overall and SSRS	1 st Grade Teacher Number of Problems and SSRS and 4 th Grade Teacher Number of Problems and SSRS

Chapter 4: Results

Assumptions

Assumptions underlying the Pearson Product Moment Correlation and the Fisher's z transformation were examined. Data were examined for linearity and although the rating scale for specific number of problems was such that the data did not meet the linearity assumption, given that the skew and kurtosis were within normal limits (± 3.0), this was not considered to be a problem. As Schatschneider and Lonigan (2010) suggest, skewed distributions may produce smaller correlations but these correlations are still valid and interpretable.

The assumption of independence was unable to be met because students were nested in teachers' classes and therefore each teacher rated several students.

Research Questions

Three research questions were posed for this study. The questions are directed at the validity of a rating scale of teachers' ratings of reading that was developed for this study. More specifically, first and fourth grade teachers rated their students' overall reading on the Reading Rating Overall Rating as well as the number of specific reading problems on the Rating of Reading Problems. The difference in the strength of the validity coefficients between first and fourth grade teachers also was compared in this study. A significance level of $p < .05$ was used when interpreting the difference in the strength of the validity coefficients. Although inclusion of statistical significance for each correlation is not included in the text, it should be noted that all correlations described as significant did attain $p < .05$ significance. The correlations with information about statistical significance are available in Tables 6-9. It is

important to note that statistical significance does not always equate to practical importance. Although all correlations described as significant did attain $p < .05$ significance, not all had practical importance. That is why only those correlations that were moderate or strong were viewed as demonstrations of validity.

The strength of the correlations is reported in the text below. The following criteria are used to describe the strength of the validity coefficients: weak validity coefficients are those below .5, moderate validity coefficients range from .5 to .6, and strong validity coefficients are .6 and above.

It is important to note that different direct measures of reading were administered to first graders versus fourth graders. Moreover, variation also existed within grades such that not all of the tests administered to first graders in the fall were administered to them in the spring. Similarly, not all of the direct measures administered to fourth graders in the fall were administered to them in the spring. Table 1 and Table 2 provide details about which tests were administered to first and fourth graders both in the fall and the spring.

Research Question 1: Are teacher ratings of first and fourth graders' overall reading performance concurrently and predictively related to direct measures of reading performance and ratings of academic competence (reading item)?

The first set of analyses addressed whether teachers' ratings on the Rating of Overall Reading were related to direct measures of reading and academic competence (reading item). Spearman Rho correlations were computed to examine the concurrent and predictive strength of the relations, among teachers' ratings of their first or fourth

grade student's on the Rating of Overall Reading and several direct assessments.

These analyses are presented in Table 6 and Table 7.

Table 6

Correlations for Question 1: Are teacher ratings of 1st overall reading performance related to direct measures of reading performance and ratings of academic competence?

1 st Grade	
Concurrent Validity	Predictive Validity
Teacher Overall and Student CTOPP ($r = .4$)	Teacher Overall and Student PRF ($r = .67$)
Teacher Overall and Student TOWRE- SWE ($r = .68$)	Teacher Overall and Student WRMT – WID ($r = .68$)
Teacher Overall and Student TOWRE – PDE ($r = .64$)	Teacher Overall and Student WRMT – Word Attack ($r = .63$)
Teacher Overall and Student WRMT – WID ($r = .67$)	Teacher Overall and Student CRAB ($r = .54$)
Teacher Overall and Student WRMT – Word Attack ($r = .61$)	
Teacher Overall and Student WIF ($r = .66$)	
Teacher Overall and Teacher SSRS ($r = .24$)	

Note. Correlations below .5 are considered weak; correlations of .5 through .59 are considered moderate; correlations of .6 and above are considered strong.

Table 7

Correlations for Question 1: Are teacher ratings of 4th graders' overall reading performance related to direct measures of reading performance and ratings of academic competence?

4 th Grade	
Concurrent Validity	Predictive Validity
Teacher Overall and Student TOWRE- SWE ($r = .52$)	Teacher Overall and Student WJ III – WID ($r = .62$)
Teacher Overall and Student TOWRE- PDE ($r = .53$)	Teacher Overall and Student WJ III – Word Attack ($r = .5$)

Teacher Overall and Student WJ III – WID ($r = .62$)	Teacher Overall and Student PRF ($r = .63$)
Teacher Overall and Student WJ III – Word Attack ($r = .49$)	Teacher Overall and Student WIF ($r = .48$)
Teacher Overall and Student PRF ($r = .65$)	Teacher Overall and Student GMRT ($r = .61$)
Teacher Overall and Student WIF ($r = .54$)	Teacher Overall and Student MAZE ($r = .65$)
Teacher Overall and Student GMRT ($r = .69$)	
Teacher Overall and Student MAZE ($r = .59$)	
Teacher Overall and Teacher SSRS ($r = .32$)	

Note. Correlations below .5 are considered weak; correlations of .5 through .59 are considered moderate; correlations of .6 and above are considered strong.

Concurrent. Teachers' ratings on the Rating of Overall Reading were significantly correlated with several direct measures of reading ability that were administered to first and fourth graders during the fall.

First grade. Students were administered five tests that assessed word recognition and decoding skills and one test that examined phonological awareness abilities. They were not administered measures of reading comprehension during the fall.

Five of the six validity coefficients were strong while the remaining validity coefficient was weak; this supports the concurrent validity of this rating scale. More specifically, teachers' ratings on the Rating of Overall Reading were strongly correlated with all five assessments of word recognition and decoding, including the TOWRE – Sight Word Efficiency ($r = .68$), the TOWRE – Phonemic Decoding Efficiency ($r = .64$), the WRMT – Word Identification ($r = .67$), the WRMT – Word Attack ($r = .61$), and the Word Identification Fluency measure ($r = .66$). In contrast,

the teachers' ratings on the Rating of Overall Reading was weakly correlated with students' performance on the CTOPP assessment of phonological awareness ($r = .4$).

Fourth grade. Fourth grade students were administered eight direct assessments in the fall, six assessing word recognition and decoding and two assessing reading comprehension. Three of the eight validity coefficients were strong, four were moderate, and one was weak. Of the six measures of word recognition and decoding, two were strongly correlated with teachers' ratings on the Rating of Overall Reading (WJ III – Word Identification: $r = .62$; Passage Reading Fluency: $r = .65$), three were moderately correlated with teachers' ratings on the Rating of Overall Reading (TOWRE – Sight Word Efficiency, $r = .52$; TOWRE – Phonemic Decoding Efficiency, $r = .53$; Word Identification Fluency, $r = .54$), and one was weakly correlated with teachers' ratings on the Rating of Overall Reading (WJ III – Word Attack, $r = .49$). In comparison, one measure of reading comprehension was strongly correlated with teachers' ratings on the Rating of Overall Reading (GMRT: $r = .69$), while another was moderately correlated with teachers' ratings on the Rating of Overall Reading (Maze: $r = .59$).

Predictive. Teachers' ratings on the Rating of Overall Reading were correlated with several direct measures of reading ability that were administered to first and fourth graders during the spring.

First grade. In the spring, first grade students were administered three measures of word recognition and decoding and one measure of reading comprehension. Three of the four validity coefficients were strong and the fourth one was moderate indicating that the measure has predictive validity in first grade. The

validity coefficients between teachers' ratings on the Rating of Overall Reading and three measures of word recognition and decoding were strong, suggesting the presence of predictive validity on this measure of teachers' ratings. These three measures included Passage Reading Fluency ($r = .67$), WRMT – Word Identification ($r = .68$), and WRMT – Word Attack ($r = .63$). Teachers' ratings on the Rating of Overall Reading was moderately correlated with the Comprehensive Reading Assessment Battery ($r = .54$) which assesses reading comprehension.

Fourth grade. In the spring, fourth grade students were administered four measures assessing word recognition and decoding and two measures assessing reading comprehension. Four of the six validity coefficients were strong, one was moderate, and one was weak. Two of the correlations between teachers' ratings on the Rating of Overall Reading and measures of word recognition and decoding were strong (WJ III – Word Identification, $r = .61$; Passage Reading Fluency, $r = .63$). Among the remaining two correlations between measures of word recognition and decoding and teachers' ratings on the Rating of Overall Reading, one was moderate (WJ III – Word Attack, $r = .5$) and the other was weak (Word ID Fluency, $r = .48$). Both measures of reading comprehension were strongly correlated with teachers' ratings on the Rating of Overall Reading (Maze, $r = .65$; and GMRT, $r = .61$).

Academic competence (reading item). Both first and fourth grade teachers completed one item on Social Skills Rating System rating their students' academic competence in reading. First grade teachers' ratings on the Rating of Overall Reading were not significantly related to their ratings' of students' academic competence in reading ($r = .24$). Similarly, fourth grade teachers' ratings on the Rating of Overall

Reading were not significantly related to their ratings of students' academic competence in reading ($r = .32$).

Research Question 2: Are teachers' ratings of the number of specific reading problems displayed by first and fourth grade children concurrently and predictively related to direct measures of reading performance and ratings of academic competence (reading item)?

The second set of analyses assessed the concurrent and predictive validity of teachers' ratings of specific number of problems in reading. The same direct assessments as described above were administered to students in the fall and spring to assess the concurrent and predictive validity of teachers' ratings. Pearson Product Moment Correlations were computed to examine the strength of the relationship between these ratings and a number of specific reading problems and several direct assessments. These analyses are presented in Tables 8 and 9. The negative correlations presented were consistent with expectations; that is, the more reading problems a student experiences, the worse he or she performs on direct measures of reading.

Table 8

Correlations for Question 2: Are teachers' ratings of the number of specific problems reading of 1st grade children related to direct measures of reading performance and ratings of academic competence?

1 st Grade	
Concurrent Validity	Predictive Validity
Teacher Number of Problems and Student CTOPP ($r = -.31$)	Teacher Number of Problems and Student PRF ($r = -.43$)

Teacher Number of Problems and Student TOWRE- SWE ($r = -.47$)	Teacher Number of Problems and Student WRMT – WID ($r = -.56$)
Teacher Number of Problems and Student TOWRE – PDE ($r = -.41$)	Teacher Number of Problems and Student WRMT – Word Attack ($r = -.50$)
Teacher Number of Problems and Student WRMT – WID ($r = -.49$)	Teacher Number of Problems and Student CRAB ($r = -.45$)
Teacher Number of Problems and Student WRMT – Word Attack ($r = -.38$)	
Teacher Number of Problems and Student WIF ($r = -.33$)	
	Teacher Number of Problems and Teacher SSRS ($r = -.25$)

Note. Correlations below .5 are considered weak; correlations of .5 through .59 are considered moderate; correlations of .6 and above are considered strong.

Table 9

Correlations for Question 2: Are teachers' ratings of the number of specific problems reading of 4th grade children related to direct measures of reading performance and ratings of academic competence?

4 th Grade	
Concurrent Validity	Predictive Validity
Teacher Number of Problems and Student TOWRE- SWE ($r = -.45$)	Teacher Number of Problems and Student WJ III – WID ($r = -.47$)
Teacher Number of Problems and Student TOWRE- PDE ($r = -.43$)	Teacher Number of Problems and Student WJ III – Word Attack ($r = -.45$)
Teacher Number of Problems and Student WJ III – WID ($r = -.50$)	Teacher Number of Problems and Student PRF ($r = -.50$)
Teacher Number of Problems and Student WJ III – Word Attack ($r = -.40$)	Teacher Number of Problems and Student WIF ($r = -.42$)
Teacher Number of Problems and Student PRF ($r = -.51$)	Teacher Number of Problems and Student GMRT ($r = -.54$)
Teacher Number of Problems and Student WIF ($r = -.49$)	Teacher Number of Problems and Student MAZE ($r = -.50$)

Teacher Number of Problems and Student
GMRT ($r = -.51$)

Teacher Number of Problems and Student
MAZE ($r = -.48$)

Teacher Number of Problems and Teacher SSRS ($r = -.28$)

Note. Correlations below .5 are considered weak; correlations of .5 through .59 are considered moderate; correlations of .6 and above are considered strong.

Concurrent. Teachers' ratings on the Ratings of Reading Problems were correlated with several direct measures of reading ability that were administered to first and fourth graders during the fall.

First grade. First grade students were administered five tests that assessed word recognition and decoding skills and one test that examined phonological awareness abilities. All five measures of word recognition and decoding were weakly and negatively correlated with teachers' ratings on Ratings of Reading Problems (TOWRE – Sight Word Efficiency, $r = -.47$; TOWRE – Phonemic Decoding Efficiency, $r = -.41$; WRMT – Word Identification, $r = -.49$; WRMT – Word Attack, $r = -.38$; and Word Identification Fluency, $r = -.33$). Additionally, the measure of phonological awareness was weakly and negatively related to teachers' ratings on Ratings of Reading Problems (CTOPP: $r = -.31$).

Fourth grade. Fourth grade students were administered eight direct assessments in the fall, six assessing word recognition and decoding and two assessing reading comprehension. Three of the eight validity coefficients were moderate while the remaining five validity coefficients were weak. This was similar to the findings with the first graders where most of the correlations between teachers' ratings on Ratings of Reading Problems and fourth graders' performance on direct

assessments were weakly and negatively related with a few moderately and negatively related. Specifically, two measures of word recognition and decoding were moderately and negatively related to teachers' ratings on Ratings of Reading Problems (WJ III – Word Identification, $r = -.50$, and Passage Reading Fluency, $r = -.51$). An additional four measures of word recognition and decoding were weakly and negatively related to teachers' ratings on Ratings of Reading Problems (TOWRE-Sight Word Efficiency, $r = -.45$; TOWRE – Phonemic Decoding Efficiency, $r = -.43$; WJ III – Word Attack, $r = -.40$; and Word Identification Fluency, $r = -.49$).

Of the two measures of reading comprehension administered to students, one was moderately and negatively related to teachers' ratings on Ratings of Reading Problems (GMRT, $r = -.51$) and one was weakly and negatively related to teachers' ratings on Ratings of Reading Problems (Maze: $r = -.48$). Although the difference in the strength of the correlations is not statistically significant, the two validity coefficients were interpreted as different, weak and moderate, because of the present study's definition of criterion for acceptable validity coefficients.

Predictive. Teachers' ratings on Ratings of Reading Problems were correlated with several direct measures of reading ability that were administered to first and fourth graders during the fall.

First grade. In the spring, first grade students were administered three measures of word recognition and decoding and one measure of reading comprehension. Two of the four validity coefficients were moderate while two of the validity coefficients were weak. Teachers' ratings on Ratings of Reading Problems were moderately and negatively correlated with two measures of word recognition

and decoding (WRMT – Word Identification, $r = -.56$; and WRMT – Word Attack, $r = -.50$) and weakly and negatively related to one measure of word recognition and decoding (Passage Reading Fluency, $r = -.47$). Additionally, the CRAB which assesses reading comprehension also was weakly and negatively related to teachers' ratings on Ratings of Reading Problems ($r = -.46$).

Fourth grade. In the spring, fourth grade students were administered four measures assessing word recognition and decoding and two measures assessing reading comprehension. Of the six validity coefficients produced, three were moderate and three were weak. One measure of word recognition and decoding was moderately and negatively related to teachers' ratings on Ratings of Reading Problems (Passage Reading Fluency, $r = -.50$), while three measures of word recognition and decoding were weakly and negatively related to teachers' ratings on Ratings of Reading Problems (WJ III – Word Identification, $r = -.47$; WJ III – Word Attack, $r = -.45$; and Word Identification Fluency, $r = -.42$). Both measures of reading comprehension were moderately and negatively related to teachers' ratings on Ratings of Reading Problems (GMRT, $r = -.54$; Maze, $r = -.50$).

Academic competence (reading item). Both first and fourth grade teachers completed the one reading item on Social Skills Rating System assessing their students' academic competence in reading. Negative correlations were expected between teachers' ratings on Ratings of Reading Problems and teachers' ratings of academic competence (reading item) because logically, the higher one's academic competence, the fewer problems one would expect that individual to experience. First grade teachers' ratings on Ratings of Reading Problems were not significantly related

to their ratings' of students' academic competence (reading item; $r = -.28$). Similarly, fourth grade teachers' ratings on Ratings of Reading Problems were not significantly related to their ratings of students' academic competence (reading item; $r = -.25$).

Research Question 3: Is there a grade-related difference in the strength of the validity coefficients?

Fisher z transformations were performed to test whether the associations between teachers' ratings of fourth graders and fourth graders' reading performance was stronger than the association between teachers' ratings of first graders and first graders' reading performance. Information about which measures were included in these analyses can be found in Table 5. Both teachers' concurrent and predictive validity were included in these analyses.

Teacher ratings of overall reading. Two of the nine validity coefficients were significantly different between first and fourth grade (Table 10). Significant differences, favoring the first grade validity coefficients, were found for teachers' ratings on Rating of Overall Reading and the TOWRE – Sight Word Efficiency, a measure of word recognition and decoding ($z = 2.73, p < .05$), and for the spring administration of WMRT – Word Attack (first grade) and WJ III – Word Attack (fourth grade; $z = 2.07, p < .05$). These findings suggest that this measure of teacher ratings on Ratings of Overall Reading has higher validity coefficients in first grade versus fourth grade for these two comparison measures of reading. There were no significant grade-related differences between teachers' ratings on the Rating of Overall Reading and the TOWRE – Phonemic Decoding Efficiency, fall administration of WMRT – Word Attack (first grade) and WJ III – Word Attack

(fourth grade), Passage Reading Fluency, WMRT – Word Identification (first grade) and WJ III – Word Identification (fourth grade), and Word Identification Fluency (Table 10).

Table 10

Correlations for Question 3: Is there a difference in the strength of the validity coefficients of Teachers Overall Ratings by grade?

Significant Differences in Teacher Overall Ratings	
Concurrent Validity	Predictive Validity
1 st Grade Teacher Overall and Student TOWRE – SWE and 4 th Grade Teacher Overall and Student TOWRE – SWE ($z = 2.73, p < .05$)	1 st Grade Teacher Overall and Student WRMT – Word Attack and 4 th Grade Teacher Overall and Student WJ III – Word Attack ($z = 2.07, p < .05$)
No Significant Differences in Teacher Overall Ratings	
Concurrent Validity	Predictive Validity
1 st Grade Teacher Overall and Student TOWRE – PDE and 4 th Grade Teacher Overall and Student TOWRE – PDE	1 st Grade Teacher Overall and Student PRF and 4 th Grade Teacher Overall and Student PRF
1 st Grade Teacher Overall and Student WIF and 4 th Grade Teacher Overall and Student WIF	1 st Grade Teacher Overall and Student WRMT – WID and 4 th Grade Teacher Overall and Student WJ III – WID
1 st Grade Teacher Overall and Student WRMT – WID and 4 th Grade Teacher Overall and Student WJ III – WID	
1 st Grade Teacher Overall and Student WRMT – Word Attack and 4 th Grade Teacher Overall and Student WJ III – Attack	

No Significant Differences in Ratings of Academic Competence

1st Grade Teacher Overall and SSRS and 4th Grade Teacher Overall and SSRS

Teacher ratings of specific reading problems. Of the nine pairs of validity coefficients that were compared, one pair was significantly different (Table 11). As hypothesized, the validity coefficient was significantly stronger for fourth grade teachers' ratings on Rating of Reading Problems and the fall administration of the Word Identification Fluency measure than for first grade teachers' ratings on Ratings of Reading Problems and the fall administration of the Word Identification Fluency measure ($z = 2.14, p < .05$). This indicates that the concurrent validity of teachers' ratings on Ratings of Reading Problems and a measure of fluency was more valid for fourth grade teachers than first grade teachers. No other significant differences were found between first and fourth grade teacher ratings on the Rating of Reading Problems and direct assessments of reading administered in the fall or the spring (Table 11).

Table 11

Correlations for Question 3: Is there a difference in the strength of the validity coefficients of Teachers' Ratings of Number of Problems by grade?

Significant Differences in Teacher Number of Problems

Concurrent Validity

1st Grade Teacher Number of Problems and Student WIF and 4th Grade Teacher Number of Problems

and Student WIF ($z = - 2.14, p < .05$)

No Significant Differences in Teacher Number of Problems

Concurrent Validity

Predictive Validity

1 st Grade Teacher Number of Problems and Student TOWRE – SWE and 4 th Grade Teacher Number of Problems and Student TOWRE– SWE	1 st Grade Teacher Number of Problems and Student PRF and 4 th Grade Teacher Number of Problems and Student PRF
1 st Grade Teacher Number of Problems and Student TOWRE – PDE and 4 th Grade Teacher Number of Problems and Student TOWRE – PDE	1 st Grade Teacher Number of Problems and Student WRMT – WID and 4 th Grade Teacher Number of Problems and Student WJ III – WID
1 st Grade Teacher Number of Problems and Student WRMT – WID and 4 th Grade Teacher Number of Problems and Student WJ III – WID	1 st Grade Teacher Number of Problems and Student WRMT – Word Attack and 4 th Grade Teacher Number of Problems and Student WJ III – Word Attack
1 st Grade Teacher Number of Problems and Student WRMT – Word Attack and 4 th Grade Teacher Number of Problems and Student WJ III – Word Attack	
No Significant Differences in Ratings of Academic Competence	
1 st Grade Teacher Number of Problems and SSRS and 4 th Grade Teacher Number of Problems and SSRS	

Academic competence (reading item). Fisher z transformations were performed to test whether the validity coefficient between teachers’ ratings of fourth graders’ reading (Ratings of Overall Reading and Rating of Reading Problems) and their ratings of academic competence (reading item) was significantly stronger than the validity coefficient between teachers’ ratings of first graders’ reading (Ratings of Overall Reading and Rating of Reading Problems) and their ratings of academic competence (reading item). There were no significant differences between teachers’

ratings on either Rating of Overall Reading or Rating of Reading Problems and their ratings of academic competence (reading item). In other words, both the concurrent and predictive validity of teachers' ratings of reading and their ratings of students' academic competence (reading item) were not significantly different in first and fourth grades.

Chapter 5: Discussion

I explored the concurrent and predictive validity of a newly created measure of teachers' ratings of students' reading behavior. This measure asked teachers to rate both students' overall reading ability as well as the number of specific reading problems experienced by students. Although direct assessments have been shown to be valid indicators of students' reading abilities, they are labor-intensive and require a large amount of time to administer (O'Connor & Jenkins, 1999). In comparison to direct assessments, teachers' ratings potentially can provide similar information in a more efficient manner.

To date, studies that have investigated the validity of teachers' ratings of students' reading performance have focused on either the concurrent *or* predictive validity of the ratings but not both. In this study teachers' ratings were correlated with assessments given during the fall as well as assessments given during the spring allowing both the concurrent and predictive validity of the ratings to be assessed. In addition, the focus of prior research has been on either general ratings of reading *or* specific reading abilities but not both. The present study used a measure in which teachers first rated their students' overall reading abilities and then identified their students' specific reading problems, thus providing ratings for both aspects. Validity of this measure of teachers' ratings also was assessed by comparing ratings of reading to teachers' ratings of academic competence (reading item).

Findings

The first hypothesis was that both first and fourth grade teachers' ratings on Ratings of Overall Reading would be moderately to strongly correlated to direct

measures of reading administered in the fall and the spring. This hypothesis was mostly supported and found to be consistent with prior research showing that teachers' ratings of overall reading performance are moderately to strongly associated with direct assessments (Hopkins et al., 1985; Kenny & Chekaluk, 1993; Greshman et al., 1987). This was shown in 21 of the 24 correlations performed for this hypothesis. Two of the correlations that did not support this hypothesis were within .02 of being classified as moderate (4th grade: fall administration of WJ III – Word Attack; spring administration of Word Identification Fluency).

The third correlation was with phonological awareness (CTOPP); it is possible that teachers are not sensitive to children's phonological awareness skills and may not take this skill into account when rating children's overall reading performance. Consistent with this suggestion, past research has identified phonological awareness as one of the areas in which teachers benefit from receiving professional development to help perfect their teaching of it (Brady et al., 2009).

The second hypothesis asserted that first and fourth grade teacher ratings on Ratings of Reading Problems would be moderately to strongly related to direct measures of reading administered in both the fall and the spring. At best, this hypothesis was weakly supported with only one third of the correlations performed being moderate and the rest being weak. Specifically, 8 of the 24 correlations performed for this hypothesis were moderate while the remaining 16 correlations performed for this hypothesis were weak.

An additional finding was that the Reading Rating Form produced higher validity coefficients for first grade teachers' ratings of word recognition and decoding

and fourth grade teachers' ratings of reading comprehension. This pattern occurred with teachers' ratings on both the Rating of Overall Reading and Ratings of Reading Problems.

Contrary to expectations, teachers' ratings on the Ratings of Overall Reading and the Ratings of Reading Problems were not significantly related to their ratings of academic competence (reading item). The present findings do not support prior research indicating that teachers' ratings of overall reading as well as teachers' ratings of the number of reading problems are both moderately to strongly related to ratings of academic competence (Feinberg & Shapiro, 2003; Hecht & Greenfield, 2001). Such differences may have resulted from only including the one reading item in the SSRS rather than the entire rating scale. Reduced variance restricts the size of correlation coefficients.

At best, findings from the present study weakly support the third hypothesis of significant grade-related differences for the validity coefficients. These expectations were based on findings by Kenny and Chekaluk (1993) and were expected because fourth graders have a wider range of skills and therefore teachers have more information available from which to rate the student. This extra information would presumably make it easier to detect a problem. In general, this measure of reading was equally valid for first and fourth grade teachers' ratings. In the three instances where there was a significant grade-related difference, higher validity coefficients were produced in first grade between teachers' ratings and measures of word recognition and decoding, and in fourth grade between teachers' ratings and fluency.

Implications

The findings from the present study suggest that teachers' ratings on the Reading Rating Form are valid measures of children's reading skills when compared with direct measures of reading. The validity coefficients are stronger for Ratings of Overall Reading than for Ratings of Reading Problems. Perhaps overall ratings are more in keeping with the way teachers normally judge their students. Alternatively, it is possible that the presence of a more restricted range in Ratings of Reading Problems versus Ratings of Overall Reading served to lower the coefficients.

Alternatively, this difference may result from differences in how teachers define reading. It is logical to posit that teachers' definitions of reading are influenced by the topics and skills emphasized in the grade they teach. Such an idea would explain why higher validity coefficients were produced for first grade teachers' ratings of word recognition and decoding and fourth grade teachers' ratings of reading comprehension. Reading comprehension likely plays a larger role in fourth grade than word recognition and decoding as the students are expected to have mastered word recognition and decoding in earlier grades. Fourth grade teachers may not focus on word recognition and decoding if they believe this should have been learned during the earlier grades. In contrast, the emphasis in first grade is placed more on word recognition and decoding than on comprehension.

For example, according to the Maryland state curriculum for reading, first grade has a total of 54 reading objectives (MSDE, 2007). Of those objectives, 18 are for reading comprehension, 13 for vocabulary, 10 for phonemic awareness, 7 for fluency, and 6 for phonics. Objectives specific to word recognition and decoding are found within the vocabulary section, the phonemic awareness section, and the

phonics section. This means that there are more first grade reading objectives dedicated to word recognition and decoding than to reading comprehension. In comparison, fourth grade has 40 reading objectives, of which 25 are for reading comprehension, 10 for vocabulary, 4 for fluency, and 1 for phonics. There is no fourth grade reading objective for phonemic awareness as students are expected to have already mastered that by fourth grade. Although skills with word recognition and decoding and reading comprehension may be important in both first and fourth grade, there are more objectives and subsequent focus dedicated to word recognition and decoding in first grade and reading comprehension in fourth grade.

Yet another explanation for the difference in teachers' ratings on the Rating of Overall Reading as compared to on the Rating of Reading Problems also pertains to the impact that teachers' definitions may have had on their ratings. Teachers only completed the Ratings of Reading Problems if their students had received a rating of a 1 or 2 on the Rating of Overall Reading. Thus, a teacher's definition of reading may have indirectly impacted whether a teacher even rated all students' number of specific reading problems.

Beyond grade-level differences, there also might be individual differences in how teachers define reading. The years a teacher has been teaching overall or teaching a specific grade as well as the highest level of education attained and completion of teaching certification may impact how one defines reading and how one rates students' reading. For example, the teacher demographic information in Table 12 shows that while some of the teachers in the present study received their teaching certification, others did not. Of those who did, the area of specialization

varied from early childhood, to special education or middle school/high school. It is possible that those who specialized in middle school/high school may have a different definition of or familiarity with the basics of reading than those who specialized in early childhood.

Table 12

Teacher Background

	1 st Grade	4 th Grade
<i>N</i>	16	17
% Female	93.8	100.0
% Certified	87.5	52.9
Total Years Teaching <i>M(SD)</i>	19.8 (13.2)	11.4 (10.4)

Limitations

There are two main limitations of the present study. The first has to do with the external validity of the results. The sample included teachers from parochial schools which were made up of primarily Caucasian students (80% Caucasian in first grade; 74% Caucasian in fourth grade). It is possible that these findings may not extend to other settings where the ethnic/racial distribution of students is different or where the teachers are working in public institutions.

A second limitation is that students were nested within teachers' classes. That is, teachers rated the students within their classes. This prevented the sample from meeting the assumption for independent data. Future research should attempt to replicate these findings by accounting for the nested nature of the data.

Implications and Future Research

One goal of the present study was to assess grade-related differences in the validity of teachers' ratings. Future studies should continue to assess whether differences exist across teachers that may impact the validity of their ratings. For example, research has yet to investigate whether teachers from different grades define reading differently. Such findings may impact the utility of teachers' ratings.

Additionally, the race of the teachers in relation to the students' race was not investigated in the present study but may be an interesting topic to study in the future. That is, future research should investigate whether the validity of teachers' ratings of reading is impacted by whether they view themselves as similar to the child, such as of the same race or same background.

As Messick (1995) suggests, a measure's validity cannot be determined without also taking into account the intended purpose of the measure. In other words, in order to determine the validity of the Reading Rating Form, the purpose and context in which it is supposed to be used needs to be considered as well. The present study assessed and supported the validity of the Reading Rating Form by correlating it with several direct assessments of reading as well as with teachers' ratings of students' academic competence (reading item). However, the present study was unable to assess whether the Reading Rating Form would meet standards of sensitivity and specificity had it been included in a screening battery. Although this measure proved to be valid, in general, within the context of the study and therefore suggests that it would be a beneficial tool to include in a screening measure, future research should continue to assess the validity of this measure by including it in a

screening battery and analyzing whether it meets standards of specificity and sensitivity.

A strength of the present study was that it included teachers' ratings of overall reading ability and ratings of specific number of reading problems in the same study thus allowing more in depth information to be retrieved than in previous studies. However, future research should go beyond studying teachers' ratings of the number of specific reading problems and investigate teachers' ratings of actual specific reading problems. A second strength of the present study was that it included a wider range of direct reading measures than included in past studies.

Although research should continue to investigate whether grade-level differences exist and whether differences between individual teachers' impact their ratings of reading, the findings of the present study support the validity of teachers' ratings on the Reading Rating Form. Given that teachers are able to provide valid ratings of reading on the Reading Rating Form, it is logical to think that with more professional development, teachers can be trained to broaden their definition of reading to include all relevant aspects beyond what is focused on in the grade they teach.

The fact that teachers in different grades may focus on different aspects of reading should not preclude the utilization of teachers' ratings. Instead, those asking teachers' to provide ratings of reading need to be cognizant of the skills emphasized in the grade level they teach and how this might impact teachers' ratings. For example, first grade teachers should be asked to provide ratings of word recognition and decoding while fourth grade teachers should be asked to provide ratings of

reading comprehension. Although these skills may be important in both grades, teachers should be asked to provide ratings using measures that have high validity coefficients. Either way, the present study indicates that teachers provide valid ratings of their students' reading abilities on the Reading Rating Form. Such findings could not come at a more opportune time when laws are mandating (NCLB, 2001) that students be proficient in reading.

Conclusion

Given these laws and the timeline they require, efficient methods of assessment would be useful for implementing universal screening and identifying children who are having difficulties in reading. Although direct assessments of reading have been shown to be valid in depicting students' skills, they take a good deal of time to complete. Universal screening is part of RTI and is the first step in identifying children who are displaying academic difficulties and who will not meet the achievement standards set forth by recent federal legislation. Existing research on teachers' ratings has looked at ratings of behavior as well as of academic abilities and specific academic abilities, such as reading. However, few studies investigating teachers' ratings of reading have established both concurrent and predictive validity of teachers' ratings of both overall reading and specific number of reading abilities all in one study. As was shown in the present study, this measure of teachers' ratings on Ratings of Overall Reading contained both concurrent and predictive validity with various direct measures assessing word recognition and decoding and reading comprehension. First and fourth grade teachers' ratings on Ratings of Reading Problems produced a greater number of weak validity coefficients compared to first

and fourth grade teachers' ratings on Ratings of Overall Reading suggesting that teachers were less able to accurately rate the specific number of students' specific reading abilities as they were their overall reading abilities. Given that this measure has demonstrated the concurrent and predictive validity of teachers' ratings in both first and fourth grades, it is hoped that it can begin to be incorporated into screenings for identifying students experiencing reading difficulties.

Appendix A

Literature Review

Concurrent Validity						
Author Name (Year)	Research Question	Students' Grade	Teacher Measure	Criterion Variable(s)	Results ^a	Comments and Notes
Hopkins, George, and Williams (1985)	1. How does the validity of teachers' ratings of reading compare with standardized tests?	4 th and 5 th grade	Investigator constructed rating of reading	Comprehensive Test of Basic Skills (CTBS, Form S, Level 2)	1. Teachers' ratings of reading performance were strongly related to the standardized reading tests (r = .73)	Teachers provided one overall rating of reading. It is unclear what these ratings were based on, what specific skills
Kenny and Chekaluk (1993)	1. Assessing the validity of 3 methods (single teacher rating, multi-item teacher questionnaire, standardized assessment) that can be used in the future to form the basis of classification for children across 3 years?	Kindergarten through 2 nd grade	Teacher Rating Scale; Teacher asked to classify reading performance into one of three categories	Lindamood Auditory; Conceptualization Test; Peabody Picture Vocabulary Test- Revised; The Syllable Counting Test ; The Word String Memory Test (WSMT) Rhyming (RHY) and Nonrhyming; Recall of Designs; Memory for	1. Teachers' responses to the questionnaire were highly correlated with the Word ID and Word Attack subtests. Teachers' ratings of older children were more valid than teachers' ratings of	Unclear if teachers' ratings of older children were more valid or if more children in the older grades were poor readers than in the younger grades

Gresham, Reschley, and Carey (1987)	1. Assess the concurrent validity of teachers' ratings of overall performance of students relative to their peers and to grade level expectations	4 th grade	Teacher Rating of Academic Performance	Sentences; Word-Identification and Word Attack subtests of Woodcock Johnson Wechsler Intelligence Scale for Children-Revised (WISC-R); Peabody Individual Achievement Test (PIAT)	younger children 1. Teachers' ratings of students relative to both peers and to grade level expectations were moderately to strongly correlated direct measures of verbal intelligence (WISC-R), reading recognition and comprehension (PIAT)
Feinberg and Shapiro (2003)	1. Are teachers' predictions of student reading performance consistent with actual oral reading fluency measured using	3 rd through 5 th grade	Reading subskills section of the Academic Competence Evaluation Scales	CBM Oral Reading Fluency	1. The teachers' ratings were shown to be moderately-strong and positively related to

	CBM?				students' oral reading fluency	
Farr and Roelke (1971)	1. Investigate the validity of various assessment procedures for measuring vocabulary, comprehension, and word analysis	5 th grade	Investigator constructed rating of these 3 skills to be completed by teacher and reading specialist	McCullough Word Analysis; Gates-McGinitie Reading Test, Survey D; California Reading Test, Elementary	1. Teachers' ratings were moderately to strongly related with standardized measures.	Teachers provided ratings on two occasions to ensure reliability; reading specialists were only available to provide ratings at one point in time
Sharpley and Edgar (1986)	1. Evaluated the accuracy of teachers' ratings of reading vocabulary, reading comprehension, and verbal intelligence when compared with standardized assessments	3 rd through 5 th grade	Investigator-designed scale comprised of three items	Progressive Achievement Tests (PAT); Peabody Picture Vocabulary Test-Revised (PPVT-R)	1. Teachers' ratings of reading vocabulary, reading comprehension, and verbal intelligence were moderately to strongly related to the direct assessments	The number of teachers participating in the study was not specified
Begeny, Eckert,	1. Examined the validity	1 st through 3 rd grade	Teacher Rating Scale	Oral Reading Fluency passages	1. The teachers'	Effect sizes of teachers'

Montarello, and Storie (2008)	teachers' ratings across 4 methods, including teachers' ratings, direct measures, teachers' estimates, and teachers' rankings		of Reading Performance; Teacher Interview Data Sheet; Class Ranking Charts	developed by Silver, Burdett, and Ginn (1991) reading series	ratings of general reading were highly, positively correlated with students' words reading fluency; strong association between teachers' estimates of instructional level and direct assessments; teachers' ranking of student's reading were strongly related to students' actual rankings	estimates were used to suggest correlations overestimate validity of teachers' ratings. However, alternative explanation could be what teachers are rating rather than analysis method
Eckert, Dunn, Coddling, Begeny, and Kleinmann (2006)	1. Investigating the extent to which teachers' perceptions of students' reading skills correspond to direct estimates of students' reading skills using two	2 nd grade	Teacher reading assessment chart; Teacher reading interview	CBM-R passages to assess oral reading fluency	1. Correlations showed moderate to strong relations between teachers' estimates of instructional level and actual instructional level.	Authors use results to indicate that correlations over-estimate validity of teachers' ratings. Alternative explanation could be that

analytic methods.

Percentage agreement analyses indicated teachers provide poor estimates of identifying instructional levels when reading in below grade level texts

the nature of what teachers are rating affects the validity rather than it only be a result of different analytical methods

Predictive Validity						
Author Name (Year)	Research Question	Students' Grade	Teacher Measure	Criterion Variable(s)	Results	Comments and Notes
Cabell, Justice, Zucker, and Kilday (2009)	1. To what extent is teacher report of children's emergent literacy skills predictive of direct behavioral assessments administered by trained assessors?	Preschool	Abbreviated 12-item version of the Clinical Evaluation of Language Fundamentals Preschool—Second Edition Pre-Literacy Rating Scale	Preschool Word and Print Awareness; Phonological Awareness Literacy Screening for Preschool	1. The teachers' ratings were moderately to positively correlated with children's print knowledge, alphabet knowledge, and writing	Emergent reading skills may differ from more advanced reading skills
Teisl, Mazzocco, and Myers (2001)	1. The purpose of the present study was to assess the	Kindergarten	Teachers' Report Form	Woodcock Johnson Psychoeducational Battery-Revised	1. Teisl et al. found a moderate correlation (<i>r</i>	Teachers provide valid ratings even of children as

	predictive value of kindergarten teachers' ratings of students for later first-grade academic achievement				=.48) between teachers' ratings of kindergarten's reading performance and direct assessment of reading a year later	young as kindergartners
Hecht and Greenfield (2001)	1. To compare the relative utility of teacher ratings versus several kinds of reading-related tests measured in the spring of first grade in predicting third grade levels of reading skills	1 st and 3 rd grade	Social Skills Rating System, Academic Competence subscale	Woodcock Johnson Tests of Achievement (Letter-Word Identification; Passage Comprehension); The Yopp-Singer Segmentation Test; Roswell-Chall Auditory Blending Test; the Stones-Concepts About Print Test; Peabody Picture Vocabulary Test - Revised	1. The amount of variance explained by teachers' ratings was not significantly different than the amount explained by all of the direct assessments, suggesting both methods are valid predictors of future reading skills	

Note. Only relevant research questions were included.

^a Correlations below .5 are considered weak; correlations of .5 through .59 are considered moderate; correlations of .6 and above are considered strong.

Appendix B

Means (Standard Deviations) of All Variables

First Grade Variable	<i>M (SD)</i>	Fourth Grade Variable	<i>M (SD)</i>
1. TOWRE – SWE (Fall)	35.97 (16.17)	1. TOWRE – SWE (Fall)	66.15 (9.31)
2. TOWRE – PDE (Fall)	16.19 (9.30)	2. TOWRE – PDE (Fall)	33.27 (10.07)
3. CTOPP (Fall)	8.43 (4.15)	3. WJ III – WIF (Fall)	52.93 (6.12)
4. WRMT – WIF (Fall)	36.51 (13.99)	4. WJ III – Word Attack (Fall)	21.83 (5.47)
5. WRMT – Word Attack (Fall)	13.42 (8.69)	5. PRF (Fall)	127.63 (27.24)
6. WIF (Fall)	24.77 (23.78)	6. WIF (Fall)	65.68 (15.44)
7. WRMT – WIF (Spring)	45.49 (11.83)	7. GMRT (Fall)	30.02 (9.53)
8. WRMT–Word Attack (Spring)	18.43 (8.95)	8. MAZE (Fall)	7.53 (2.42)
9. PRF (Spring)	65.66 (38.44)	9. WJ III – WIF (Spring)	54.69 (5.78)
10. CRAB (Spring)	4.09 (2.30)	10. WJ III – Word Attack (Spring)	23.58 (4.59)
11. Rating of Overall Reading	3.25 (.96)	11. WIF (Spring)	77.44 (17.13)
12. Rating Number of Problems	.62 (1.39)	12. GMRT (Spring)	32.82 (10.24)
13. Ratings of Academic Competence	3.54 (1.13)	13. MAZE (Spring)	8.95 (2.60)
		14. PRF (Spring)	138.96 (39.28)
		15. Rating of Overall Reading	3.23 (.92)
		16. Rating Number of Problems	.62 (1.39)

References

- Al Otaiba, S., Petscher, Y., Pappamihiel, N., Williams, R., Dyrland, A., & Connor, C. (2009). Modeling oral reading fluency development in Latino students: A longitudinal study across second and third grade. *Journal of Educational Psychology, 101*, 313-329.
- Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.
- Batsche, G., Elliot, J., Graden, J., Grimes, J., Kovaleski, J., Prasse, D., et al. (2006). Response to intervention: Policy considerations and implementation. Alexandria, VA: National Association of State Directors of Special Education.
- Begeny, J., Eckert, T., Montarello, S., & Storie, M. (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.
- Brady, S., Gillis, M., Smith, T., Lavalette, M., Liss-Bronstein, L., Lowe, E., North, W., Russo, E., and Wilder, T. D. (2009). First grade teachers' knowledge of phonological awareness and code concepts: Examining gains from an intensive form of professional development and corresponding teacher attitudes. *Reading and Writing: An Interdisciplinary Journal, 22*, 425-455.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/ Correlation Analysis for the Behavioral Sciences* (3rd Ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
- Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryan, J. D. (2006). Selecting at-risk

- readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology*, 98, 394-409.
- Demaray, M. K., & Elliott, S. N. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. *School Psychology Quarterly*, 13, 8-24.
- DiPerna, J. C., & Elliott, S. N. (1999). Development and validation of the academic competence evaluation scales. *Journal of Psychoeducational Assessment*, 17, 207-225.
- Donovan, M. S., & Cross, C. T. (2002). *Minority students in special and gifted education*. Washington, DC: National Academy Press.
- Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test—Revised*. Circle Pines, MN: American Guidance Service.
- Dupaul, G. J. & Rapport, M. D. (1991). Teacher ratings of academic skills: The development of the academic performance rating scale. *School Psychology Review*, 20, 284-300.
- Fisher, R. A. (1970), *Statistical Methods for Research Workers*, Fourteenth Edition, Davien, CT: Hafner Publishing Company.
- Fuchs, L. S. (n.d.). *Project PROACT Maze Reading Passages*. Nashville, TN: Vanderbilt University.
- Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review*, 21, 45-58.
- Fuchs, L.S., Fuchs, D. & Compton, D.L. (2004). Monitoring early reading

development in first grade: Word identification fluency versus nonsense word fluency. *Exceptional Children*. 71, 7-21.

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

Guthrie, J. T., Siefert, M., Burnham, N. A, & Caplan, R.I. (1974). The maze technique to assess, monitor reading comprehension. *The Reading Teacher*, 28 161-168.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied Statistics for the Behavioral Sciences* (5th ed.). NY: Houghton Mifflin Company.

Individuals with Disabilities Education Improvement Act of 2004, 20 U.S.C. § 1400 et seq. (2004)

Lee, J., Grigg, W., and Donahue, P. (2007). *The Nation's Report Card: Reading 2007* (NCES 2007-496). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). *Gates-MacGinitie Reading Tests, Fourth Edition, Forms S and T*. Itasca, IL: Riverside Publishing.

Maryland State Department of Education. (2007). Using the State Curriculum: Reading/ELA, Grade 4. Retrieved from

<http://mdk12.org/instruction/curriculum/reading/standard1/grade4.html>

Maryland State Department of Education. (2007). Using the State Curriculum: Reading/ELA, Grade 1. Retrieved from

<http://mdk12.org/instruction/curriculum/reading/standard1/grade1.html>

- Messick, S. (1995). Validity of psychology assessment: Validation of interferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741-749.
- No Child Left Behind Act of 2001, 20 U.S.C. § 6319 (2008).
- O'Connor, R. E., & Jenkins, J. R. (1999). The prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading*, *3*, 159-197.
- Roswell, F. G. & Chall, J. S. (1959). Roswell-Chall Auditory Blending Test. NY: Essay Press.
- Schatschneider, C., & Lonigan, C. J. (2010). Misunderstood statistical assumptions undermine criticism of the National Early Literacy Panel's report. *Educational Researcher*, *39*, 347-351.
- 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.
- Tabachnick, B. G. & Fidell, L. S. (2001). *Using Multivariate Statistics* (4th ed.). Needham Heights, MA: Allyn and Bacon.
- Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of Word Reading Efficiency: Examiner's Manual*. Austin, TX: Pro-Ed.
- Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *Comprehensive Test of Phonological Processing: Examiner Manual*. Austin, TX: Pro-Ed.
- Wiig, E. H., Secord, W. A., & Semel, E. (2004). *Clinical Evaluation of Language Fundamentals Preschool—Second Edition*. San Antonio, TX: Harcourt Assessment.

- Wood, J. M., Garb, H. N., & Nezworski, M. T. (2007). Psychometrics: Better measurement makes better clinicians. In S.O. Lilienfeld & W.T. O'Donohue (Eds.), *The great ideas of clinical science* (pp.77-92). NY: Routledge.
- Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Tests of Achievement (Standard Battery, Form B)*. Allen, TX: DLM Teaching Resources.
- Woodcock, R., McGrew, K., Mather, N., & Schrank, F. (2001). *Woodcock- Johnson Tests of Achievement, Third Edition*. Itasca, IL: Riverside Publishing.
- Yopp, H. K. (1988). The validity and reliability of phonemic awareness tests. *Reading Research Quarterly*, 23, 159-177.
- Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science