1-1-2006

Incremental and predictive utility of formative assessment methods of reading comprehension.

Amanda M. Marcotte
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_1

Recommended Citation

Marcotte, Amanda M., "Incremental and predictive utility of formative assessment methods of reading comprehension." (2006). *Doctoral Dissertations 1896 - February 2014*. 5765.
https://scholarworks.umass.edu/dissertations_1/5765

This Open Access Dissertation is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations 1896 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

★ UMass/AMHERST ★

312066 0324 9868 1

LD
3234
M267
2006
M3213

University of
Massachusetts
Amherst

L I B R A R Y

[illegible]

HIGHSMITH #45115

This is an authorized facsimile, made from the microfilm master copy of the original dissertation or master thesis published by UMI.

The bibliographic information for this thesis is contained in UMI's Dissertation Abstracts database, the only central source for accessing almost every doctoral dissertation accepted in North America since 1861.

UMI[®] Dissertation
Services

From: ProQuest
COMPANY

300 North Zeeb Road
P.O. Box 1346
Ann Arbor, Michigan 48106-1346 USA

800.521.0600 734.761.4700
web www.il.proquest.com

Printed in 2007 by digital xerographic process
on acid-free paper

 

INCREMENTAL AND PREDICTIVE UTILITY OF FORMATIVE
ASSESSMENT METHODS OF READING COMPREHENSION

A Dissertation Presented

by

AMANDA M. MARCOTTE

Submitted to the Graduate School of the University of Massachusetts Amherst in
partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

MAY 2006

STUDENT DEVELOPMENT AND PUPIL PERSONNEL SERVICES
DEPARTMENT OF THE SCHOOL OF EDUCATION

SCHOOL PSYCHOLOGY PROGRAM

INCREMENTAL AND PREDICTIVE UTILITY OF FORMATIVE
ASSESSMENT METHODS OF READING COMPREHENSION

A Dissertation Presented

by

AMANDA M. MARCOTTE

Approved as to style and content by:

John M. Hintze, Chair

William J. Matthews, Member

James M. Royer, Member

Christine McCormick, Department Head
Dean of the School of Education

INCREMENTAL AND PREDICTIVE UTILITY OF FORMATIVE
ASSESSMENT METHODS OF READING COMPREHENSION

A Dissertation Presented

by

AMANDA M. MARCOTTE

Approved as to style and content by:

John M. Hintze, Chair

William J. Matthews, Member

James M. Royer, Member

Christine B. McCormick, Dean
School of Education

DEDICATION

I would like to dedicate this dissertation to my family and my friends, those who inspire me to relish in life at every turn. Those who laugh often and find the greatest joys in the simplest of things. Your love provided me a place from where to start and your support the light upon the path.

I am dedicating this dissertation to my dad who always taught me to aim for the moon thus fall among the stars; to my mother who inspired my love of learning and a drive to help others attain the value in education; to my sister for always carrying her "sister purse" filled with more support than merely the obvious; and to my nephew whose laughter unfurls on the wind as should the laughter of all children for whom we serve. Finally, this dissertation is dedicated to my grandparents for their unwavering support. I am who I am because of their love.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to each member of the school psychology faculty, Gary Stoner, John Hintze and Bill Matthews, as well as to Mike Royer for serving on my dissertation committee. First, I would like to extend my gratitude to Mike Royer for his professional support and his constructive suggestions toward the theoretical foundations of this research project. I would like to thank Bill Matthews for his energetic encouragement throughout the process of this research and his ability to weave humor into my professional development over the years. To Gary Stoner, I would like to extend my deepest gratitude for modeling a patient, empathetic and responsible leader and professional, and encouraging me to hold sacred the notion of service to the field of school psychology.

I would like to extend a very special thank you to my advisor and Committee Chair, John Hintze, for his relentless faith in me and patient and thoughtful guidance over the last five years. I am so grateful for his rational support for each endeavor I pursued throughout my graduate school career. There are no words that can accurately express my sincere gratitude for his advisorship, mentorship and friendship.

Thank you to Stan Scarpati, Chris Parker and Judy Loughlin for providing me diverse professional experiences throughout these years that have allowed me to develop well-rounded skills beyond all I learned in the classroom. Thank you to all the students who helped me with my research, especially Pam Plumer, Maren Watkins and Julie Clarfield, for being my support and my family every step of the way. Finally, I would like to extend a special thank you to the principals, Reading Specialists, Fourth Grade Teachers and to all the students who participated in the process of my data collection after a long season of testing. Their willingness to participate made this project happen.

ABSTRACT

INCREMENTAL AND PREDICTIVE UTILITY OF FORMATIVE ASSESSMENT METHODS OF READING COMPREHENSION

MAY 2006

AMANDA M. MARCOTTE, B.A, PROVIDENCE COLLEGE

M.Ed., UNIVERSITY OF MASSASHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor John M. Hintze, Ph.D.

Formative assessment measures are commonly used in schools to assess early literacy skills as indicators of reading acquisition and to design instruction accordingly. The purpose of this research was to investigate the incremental predictive validity of formative assessment measures of reading comprehension. It was hypothesized that formative measures of reading comprehension will contribute more to our understanding of students' overall reading abilities than simply Oral Reading Fluency (ORF). It was also hypothesized that these measures can be modeled in a meaningful way to explain student performance on criterion measures of academic competence. Four formative measures of reading comprehension – Maze (MZ), Retell Fluency (RTF), Written Retell (WRT), and Sentence Verification Technique (SVT) – were used to measure unique aspects of reading comprehension through production-type responses in an efficient and instructionally meaningful way. The Group Reading Assessment and Diagnostic Evaluation (GRADE) was used as a measure of overall reading proficiency while the Massachusetts Comprehensive Assessment System (MCAS) was used as a measure of academic competence. Data were collected from 111 fourth grade students from two Western Massachusetts elementary schools. Four multiple regression equations were

computed to test the hypothesis that measures of reading comprehension will contribute more to our understanding of reading proficiency than ORF alone. Each measure significantly predicted performance on GRADE above and beyond that predicted by ORF. A larger multiple regression equation was used to evaluate which measure predicted a unique and significant proportion of the variance in reading proficiency. The MZ, SVT and WRT were significant predictors in the model. Using the measures found to be significant with ORF, a logistic regression analysis was computed to evaluate how reliably the newly constructed model predicted a pass or fail status on the fourth grade English Language Arts section of MCAS. None of the predictors were significant. An additional multiple regression analysis was computed to predict MCAS performance and the newly constructed model of reading proficiency. This model predicted approximately .66 of the variability in MCAS performance.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	v
ABSTRACT	vi
LIST OF TABLES	xi
LIST OF FIGURES	xii
 CHAPTER	
1. INTRODUCTION TO THE RESEARCH: BACKGROUND AND PURPOSE	1
Chapter Summary.....	1
Federal Call for Improved Literacy Instruction.....	2
Systematic Formative Evaluation and Curriculum-based Measurement.....	4
Formative Assessment of Basic Early Literacy Skills.....	6
Oral Reading Fluency as an Indicator of Overall Reading Competence.....	7
Research Agenda for Improving Reading Comprehension Outcomes.....	12
Formative Assessment Measures of Reading Comprehension....	13
Purpose of this Study.....	18
Research Questions and Hypotheses.....	19
 2. REVIEW OF THE LITERATURE: THE RESEARCH ON READING COMPREHENSION INSTRUCTION AND ASSESSMENT	 21
The Current State of Literacy in America	21
Literacy Instruction and the Federal Government	22
Scientificallly-based Educational Research” and Practice	24
Convergent Evidence about Literacy Instruction	26
The State of the Research Regarding Reading Comprehension ...	28
Evidence in Reading Comprehension Instruction	31
Enabling Factors	33
Cognitive Strategies and Teaching Approaches	34
Summarizing the Evidence in Reading Comprehension Instruction	49
Assessing Reading Comprehension.....	53

Implications for Practice	60
3. METHOD	62
Participants	62
Settings	62
Materials	63
Independent Variables	63
Individually Administered Measures.....	63
Group Administered Measures	64
Criterion Measures	67
Procedures	70
Training Data Collectors	70
Interrater Agreement and Procedural Integrity	71
Data Collection	71
Data Analysis	73
Screening	73
Examining Relationships between Variables	73
Exploring Variability in Reading Comprehension	74
Deriving a Model of Reading Ability	74
Predicting Academic Achievement	75
4. RESULTS OF THE ANALYSIS	76
Summary of Purpose and Method	76
Screening for Underlying Assumptions	76
Interrater Agreement and Procedural Integrity	78
Examining Relationships between Variables	79
Exploring Variability in Reading Comprehension	82
Deriving a Model of Reading Ability	83
Predicting Academic Achievement	84
5. DISCUSSION OF RESULTS	86
Summary of the Research	86
Results of the Analysis	87
Limitations of this Study	90
Implications for Practice	91
Directions for Future Research	96

APPENDICES

A. SUMMARY OF DEPENDENT VARIABLES FROM THE NRP

STUDIES.....	100
B. CONSENT FORM	101
C. ASSESSMENT INTEGRITY CHECKLIST	104
D. Q-Q PLOTS OF UNIVARIATE DISTRIBUTIONS	105
REFERENCES	108

LIST OF TABLES

Table	Page
3.1: The counterbalancing design of group administered tests	72
4.1: Descriptive statistics of dependent and independent variables	78
4.2: Correlation matrix for variables in the study	81
4.3: Multiple regression analyses of ORF with measures of reading comprehension ..	82
4.4: Overall regression model for reading proficiency $GRADE_i = \beta_0 + \beta_1 (ORF_i) + \beta_2 (MAZE_i) + \beta_3 (WRT_i) + \beta_4 (RTF_i) + \beta_5 (SVT_i) + \varepsilon_i$	83
4.5: Stepwise comparison of independent variables	84
4.6: Results of logistic regression analysis	84
4.7: Overall regression model for academic achievement $MCAS_i = \beta_0 + \beta_1 (ORF_i) + \beta_1 (MAZE_i) + \beta_2 (SVT_i) + \beta_3 (WRT_i) + \varepsilon_i$	85

LIST OF FIGURES

Figure	Page
4.1: Scatterplots to examine interrater variability of ORF and RTF.....	79

CHAPTER 1

INTRODUCTION TO THE RESEARCH: THE BACKGROUND AND PURPOSE

Chapter Summary

To be successful in today's society, students must have greater skills than ever before. American schools are filled with students from increasingly more diverse backgrounds, and the demand upon teachers to bring every student to a higher level of proficiency requires the use of the most efficacious instructional practices. This chapter describes how the five critical elements of effective literacy instruction - phonemic awareness, phonics, fluency, vocabulary, and comprehension - were articulated by the National Reading Panel (2000), and how the use of formative assessment strategies in a targeted and systematic way helps guide instruction of these skills for developing readers.

Simple and valid production-based assessment tools are readily available to measure reading enabling skills: phonemic awareness, phonics and fluency. These tools allow educators to frequently assess their students' abilities on important indicators that are predictive of reading success and modify their instruction as necessary. Oral reading fluency is most often used as an indicator for overall reading competence, but research suggests it is less predictive of academic success for older students who have gained a level of reading competence in which they read with speed, accuracy and prosody (Shinn, Good, Knutson, Tilly & Collins, 1992). At the point when decoding is automatized for students and proficient reading can no longer be inferred from reading rate alone, production-based assessment tools are needed that allow educators to observe and quantify indicators of reading comprehension development, similar to how we can currently observe the development of the enabling skills. Currently, there exist some tools

for measuring aspects of reading comprehension that may be useful indicators of reading comprehension skill development. This chapter describes four of these tools: maze, oral retell, written retell and sentence verification technique, and how they might contribute to our understanding of reading proficiency beyond information gained by oral reading fluency.

Federal Call for Improved Literacy Instruction

In response to the 1997 congressional charge, the National Reading Panel (NRP) was established to investigate the present state of research regarding effective strategies to teach children to read. Using a rigorous selection process, the Panel focused on experimental and quasi-experimental studies to address the question whether instruction in five specified areas improves reading achievement. Their investigation was designed to uncover the critical features of effective literacy instruction, and identify effective strategies for teaching the essential skills.

The NRP identified five essential components of literacy instruction: alphabetic awareness including both phonemic awareness and phonics, strategies to increase reading fluency, vocabulary instruction and comprehension strategy instruction. Results of their findings suggested instruction in phonemic awareness showed impressive results in the acquisition of reading and spelling. Systematic and explicit phonics instruction - teaching letter-sound correspondences in a direct and targeted sequence - appeared to be an essential feature in a successful reading program. The goal of such instruction is to increase fluent and accurate word recognition skills. In addition, building reading fluency was also found to be an important foundational skill to quality literacy instruction. Moreover, reading fluency was found to be critically related to comprehension such that

children who have adequate reading fluency skills are more likely to demonstrate adequate reading comprehension skills as compared to those students whose fluency skills are deficient. Although the Panel found scarce evidence regarding the best instructional methods for teaching vocabulary, they did find that vocabulary instruction increases students' comprehension skills.

Finally, the Panel targeted text comprehension as an additional critical feature of effective literacy instruction. The Panel adopted the definition of reading comprehension from Harris and Hodges' 1995 Literacy Dictionary which described comprehension "intentional thinking during which meaning is constructed through interactions between text and reader." Of the 16 types of text comprehension instruction, the NRP found seven strategies to have solid scientific support in enhancing students' abilities to comprehend as measured by various experimenter-designed comprehension tests and some standard tests. Of these strategies four are cognitive strategies to aid students' understanding of text, comprehension monitoring, question answering, question generation, and summarization. They found two effective strategies for teaching comprehension to students, cooperative learning and reciprocal teaching strategies, and graphic and semantic organizers. Importantly, they found teaching a combination of strategies most advantageous to improving text comprehension.

Interestingly, while numerous programs rigorously follow the NRP framework, differences among curricula abound with some offering more instruction in some critical areas than others offer. Many curriculum packages provide an overabundance of material that teachers must prioritize in their classes. Most lack a specific sequence and/or ordering of skills as they should be presented to students leaving teachers to determine

the balance in what they teach. This, combined with differing student needs, presents a challenge to classroom teachers as they attempt to teach reading comprehension systematically and strategically. Although the NRP recommends the importance of , assessing the needs of individual students and to tailor instruction accordingly, this cursory recommendation is laden with questions regarding what tools are available to teachers to conduct such invaluable assessments to sculpt their instruction accordingly.

Systematic Formative Evaluation and Curriculum-based Measurement

Systematic formative evaluation is an inductive approach to individualized instruction where a student's academic progress is regularly monitored under varying instructional methods. Frequent and systematic evaluation allows educators to modify instruction as necessary until adequate student progress is determined according to the student's learning rates as indicated by slope changes among the data points. Thus, it is a functional process through which educators are able to observe changes in students' rates of learning in relation to instructional procedures.

In 1986, Fuchs and Fuchs conducted a meta-analysis on 21 studies that operationalized systematic formative evaluation as "curriculum-based data collection that occurred at least twice each week, with decisions concerning the adequacy of programs formulated on an individual, not group, basis (p.200)." They found an effect size of .73, demonstrating the use of curriculum-based measures to inform instructional decisions can significantly improve the academic outcomes of students with mild learning disabilities. Additionally, they found even greater effect sizes when the data were graphed, making it possible to observe changing rates in progress compared to when the data were simply

recorded descriptively. Thus research demonstrates that the use of systematic formative evaluation can significantly improve academic performance for struggling students.

Such evaluation depends upon the availability of valid and reliable measurements to assess specific academic skills. Curriculum-based measurement (CBM) procedures allow educators to quantify critical academic skills via direct observations of those skills. Research has demonstrated that CBM elicits reliable and valid data describing students' current skills (Marston, 1989). These measures are also extremely useful in illustrating progress on specific skills over time.

An innovation at the time, CBM was designed to combine traditional psychometric principles with subjective teacher observational judgements (Deno, 1985). The goal was to design simple and efficient assessment techniques that directly reflected the classroom curriculum and thus bridge assessment results to instructional planning. The assessments were designed to be inexpensive and exist in many forms to be given frequently so as to assess student progress in relation to changes in instruction. Finally, the assessment results are simple to understand and thus communicate with ease students' skill level and progress to teachers, parents and administrators. CBM has proven useful in these goals and has also evolved into an accountability tool for educators to demonstrate student's response to progressively more intensive interventions.

In addition to the critical elements delineated by the NRP, systematic formative evaluation, using valid and reliable measures that directly observe specific academic skills, is an essential feature of quality literacy instruction. The utilization of such practice, in concert with teaching strategies that have been demonstrated effective, allow

educators to work from a prevention framework where they are able to identify struggling students and teach directly to their needs to remediate the problem early and effectively.

Formative Assessment of Basic Early Literacy Skills

Currently there exist formative assessment measures that allow educators to observe and quantify early literacy skills in the areas of phonemic awareness, letter-sound correspondences and oral reading fluency. Because of the stark and multiplicative nature of reading failure, such tools were designed to identify children struggling to learn at an early age and thus remediate the problem as soon as possible. Evidence indicating phonemic awareness and agility with the alphabetic code is predictive of reading success, along with evidence that instruction in these basic skills can prevent reading failure catalyzed efforts in the design of early literacy assessments and instructional strategies (Adams, 1990). For example the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) were designed to directly assess these predictive skills and to identify students who are at risk for reading failure and remediate the problem through targeted early literacy instruction.

Kaminski and Good (1989) wrote tools assessing early academic skills must minimally:

- a.) document and account for growth on a continuum of foundational reading skills,
- b.) predict success or failure on criterion measures of performance (i.e., high stakes tests), and
- c.) provide an instructional goal that if met will prevent
- d.) reading failure and promote reading success. (p.260)

For example the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) were designed to directly assess skills predictive of later reading success so as to identify students who are at risk for reading failure as early possible as and remedy the problem

through targeted early literacy instruction. A few of the screening assessments within the DIBELS battery are directly related to critical elements of early literacy instruction and sensitive to student growth as a function of instructional changes. Phonemic awareness is assessed using two of the tests, Initial Sound Fluency (ISF) and Phoneme Segmentation. Phonics and knowledge of letter-sound correspondences are efficiently measured using Letter Naming Fluency (LNF) and Nonsense Word Fluency (NWF). These brief assessments, and others including Letter Sound Fluency and Word List Reading, have both construct validity and instructional utility to identify where decoding skills breakdown. Thus assessments, such as those contained within the DIBELS, provide teachers the ability to assess their students' most basic decoding skills, assign instructional groups accordingly, monitor progress and modify instruction as necessary.

Unfortunately, the utility of DIBELS and other indicators of early reading skills for guiding literacy instruction ends at the decoding subcomponents of reading. Most educators would agree the primary goal of reading is comprehension. Thus there is still a need for equally informative tools to assess reading comprehension that also account for growth on a continuum of developing skills, predict success or failure on criterion measures of performance and provide meaningful instructional goals.

Oral Reading Fluency as an Indicator of Overall Reading Competence

“Fluent reading is intrinsically elegant in both form and cadence. We certainly know it when we see it, and we are quick to celebrate it, along with the trajectory of success it portends.”(Kame enui & Simmons, 2001, p.)

While the fundamental goal of reading may be to extract meaning from written words, fluent reading in itself is a laudable accomplishment. It is a complex orchestration of many cognitive activities, each requiring a level of automaticity so as to effectively

execute in concert with one another. Fluent reading necessitates both lower-order subskills involved in word level construction and higher-order processes necessary for meaning construction. It requires the ability to recognize printed words efficiently while nearly immediately calling upon the meaning of that word and eventually compiling the meaning of the entire text.

Rapid word recognition consists of the subcomponent skills of phonological segmenting and recoding (Fuchs, Fuchs, Hosp & Jenkins, 2001). Initially the reader must utilize phonological agility to recognize letter sounds in concert with perceptual agility to identify letter-sound correspondences as coded through the alphabetic principles. Eventually, the reader then recognizes the entire lexical structure by compiling the letter representations and their corresponding sounds into whole word units. Once the word is read as an entire unit, the recoded lexical structure is immediately processed for meaning construction.

Theoretically, the lexical input is mapped upon prior knowledge to access both the word and its meaning from stored memory. Kintsch (1988) describes this process through a construction-integration model in which the reader constructs a text base from accessing previous knowledge from linguistic input. The text construction phase continues to activate prior knowledge with each lexical input, while simultaneously the information is integrated until a stable and meaningful interpretation of the text can be established.

Since the inception of the LaBerge and Samuels (1974) automaticity theory, reading theorists have constructed their own ideas of reading competence upon the assumption that we have limited attentional capacity and thus complex cognitive

processes, such as reading, require the automation of subcomponent skills to effectively execute. The automaticity theory purports that attentional capacity is limited, thus the subcomponent skills of any complex cognitive task must be automatized so as to unencumber attentional capacity. Thus, reading tasks require word recognition processes to be automatized because cognitive processes involved in meaning construction demand large amounts of cognitive attention.

Reading fluency has been ascertained the most salient characteristic of skillful reading (Adams, 1990). It theoretically encompasses all the subcomponents of reading. Conveniently, it can be directly observed through read aloud tasks and, thus, plays a pivotal role in the assessment of reading skills. When quantified, it is arguably an excellent overall measure for the subcomponent skills involved in decoding: phonemic awareness skills such as segmenting and decoding and the application of the alphabetic principle necessary for blending and recoding. Fuchs, et al. (2001) describe oral reading fluency (ORF) as “a direct measure of phonological segmentation and recoding as well as rapid word recognition (p.239)” as they presented evidence identifying it as a reliable indicator of overall reading competence.

In one study, researchers contrasted four measures of reading comprehension - a question and answer test, an oral retell task, a cloze assessment and oral reading - with a criterion measure. They found that oral reading fluency more closely correlated with the criterion measures than did the three other measures (Fuchs, Fuch & Maxwell, 1988). Jenkins, Fuchs, Van den Broek, Espin and Deno (2003) conducted another study supporting the hypothesis that ORF is a good indicator of overall reading competence. Here, the authors explored the relationship between list reading and context reading.

Results suggested that comprehension skills play a role in oral reading fluency by demonstrating that context reading fluency is a stronger predictor of reading comprehension than list reading fluency. These researchers concluded fluent text reading reflects both rapid decoding as manifested in facile word identification skills and comprehension processes.

Amid the current education reform movement where state and local school administrators depend upon the results of high-stakes performance tests to evaluate the evolution of their educational programs and resource allocations, teachers must ensure their students succeed in the face of such performance indicators. Assessment tools with the capability of predicting each student's performance on such high stakes tests offer teachers useful information allowing them to strategically modify and intensify their instruction to prepare their students accordingly. ORF has been found to be a reliable tool for such instructional planning (Good, Simmons & Kame'enui 2001). Subsequently, various studies have explored the relationship between oral reading fluency and specific statewide assessments, confirming third grade fluency scores are excellent predictors of performance on a variety of state performance tests (Barger, 2003, Buck & Torgesen, 2003 & Shaw & Shaw, 2002,). These researchers have also been able to identify reliable cut-off scores of oral reading rates that provide educational utility and meaningful goal setting.

Interestingly, research indicates ORF, as an indicator of overall reading competence, has significant predictive power for students in third grade and early fourth grade, whereas data points collected later in grade four appear less predictive of student performance on high-stakes criterion-referenced tests (Stage & Jacobson, 2001). One

explanation for the limited utility of fourth grade fluency scores is the fact in earlier grades students' growth rates in words read correctly per minute are significantly larger than the gains made by students in fourth grade (Fuchs, Fuchs, Hamlett, Waltz & Germann, 1993). Thus the reduced predictive power might be the result of less variability between individual fluency rates. While oral reading fluency may reach a ceiling effect, it is possible that there are additional indicators of reading competence in addition to the single oral reading measure.

In another study, Shinn et al. (1992) further explored the variables subsumed within the construct of fluent reading. Specifically, the authors questioned whether ORF was a measure of overall reading competence, simply decoding, an assessment of comprehension or a distinct feature to itself. They wrote:

Given the primary importance of comprehension in the reading process, caution would be indicated when making other decisions about reading skills based only on decoding skills. In contrast, if oral reading fluency provided an adequate measure of comprehension skills, its robustness in educational decision-making would be supported (p. 461).

Using confirmatory factor analysis, Shinn and colleagues examined the theoretical factor structures of reading fluency by comparing four models: (a) a unitary model where decoding and fluency were not distinct from one another, (b) a two-factor model in which fluency was subsumed into decoding and comprehension was a distinct factor, (c) a two-factor model consisting of decoding and comprehension, where fluency was subsumed into comprehension, and (d) a three-factor model in which decoding, fluency and comprehension were three distinct factors.

Results provided additional support that oral reading fluency is in fact a good measure of reading proficiency and comprehension; however, this relationship varied

developmentally. In particular, the model that best fit reading fluency for third grade students was the single factor model in which decoding and comprehension were subsumed under one construct and not distinct from one another. Interestingly, the two-factor model in which fluency and decoding represented one factor and comprehension a distinct entity best fit the fifth grade data. This suggested once decoding is mastered, oral reading fluency no longer accounts for the variability in reading proficiency without considering the unique role comprehension plays for the more proficient reader.

As decoding becomes fully automatized and oral reading fluency becomes less useful in discriminating individual differences in reading ability, there is a need for equally useful assessment tools that can account for the variability in reading achievement based on reading comprehension skills. Like their predecessors such tools must be useful formatively - sensitive to academic growth and instructional changes - and equally predictive of academic success as reading fluency is for developing readers.

Research Agenda for Improving Reading Comprehension Outcomes

In 1999, the Office of Educational Research and Improvement of the U.S. Department of Education charged the RAND reading study group with the task of developing a research agenda to address literacy instruction. The study group prioritized the activities of their agenda beyond the early literacy developments of the previous decade and focused on instruction, teacher training and assessment of reading comprehension. While the group identified instructional practices that show promise to improve comprehension skills, research evaluating the effectiveness of such programs is dependent upon valid and reliable measures of reading comprehension. As a result, the study group prioritized research in the area of comprehension in their agenda. They

wrote, “Teachers who are interested in improving their instruction need reliable and valid assessments that are closely tied to their curriculum so they can identify those students who are learning and those who need extra help” (Executive Summary, xix).

One major recommendation of the RAND report was a call for the exploration of measures of comprehension that reflect the dynamic and developmental nature of reading comprehension. In particular, the study group listed ten minimal features of such assessment:

1. The capacity to reflect authentic outcomes;
2. Congruence between assessments and the processes involved in comprehension;
3. Developmental sensitivity;
4. The capacity to identify students as poor comprehenders;
5. The capacity to identify subtypes of poor comprehenders;
6. Instructional sensitivity;
7. Openness to intra-individual differences;
8. Usefulness in instructional decision-making;
9. Adaptability with respect to individual, social, linguistic and cultural variation; and
10. A foundation in theory and psychometrics (pp. 56-57.)

Formative Assessment Measures of Reading Comprehension

Most would agree reading comprehension is a complex and reciprocal process involving the interaction between word identification processes, the integration of prior knowledge, vocabulary and language knowledge, and cognitive monitoring strategies (Adams, 1990). Howell and Nolet (2000) categorize these processes into enabling skills and comprehension strategies. They describe comprehension development as procedural learning in which successful readers have the necessary enabling skills that allow them to employ effective comprehension strategies and thus be active readers. Howell and Nolet describe the assessment of reading comprehension as an evaluative process, which hierarchically rules out enabling skills, such as decoding, vocabulary, syntactic awareness

and prior knowledge. Subsequently the evaluation explores how the student employs comprehension strategies, including monitoring meaning, selective attention to the reading task, connecting text to prior knowledge and clarification of confusions.

Because reading comprehension is interactive by nature, it is a nebulous construct to observe directly and assess. The evaluative process described by Howell and Nolet includes the use of structured interviews and observations of specific tasks, such as observing students adjust their reading rate in accordance to text difficulty and rereading confusing sections in a text. Such active reading strategies are difficult to quantify. Yet currently there are measures that show promise for the evaluation of specific skills and strategies that play important roles in the comprehension process. It is necessary to explore these quantitative assessment tools for their construct validity and treatment utility as indicators of overall reading comprehension ability.

In response to the call formative measurements be efficient to administer and created in multiple forms for continuous assessment, maze techniques have been employed as a viable option for the assessment of reading comprehension (Deno, 1985). The maze (MZ) task consists of a grade-level reading passage in which every *n*th word is deleted and substituted with three word choice options. Students must read the passages and circle the word that best fits the meaning of the sentence. Some theorists criticize the maze task merely measures sentence meaning and structural repetition and answers do not depend upon the context in which the sentence exists (Parker, Hasbrouck & Tindal, 1992), while others believe it is a meaningful task by which reading comprehension can be measured (Guthrie, Siefert, Burnham & Caplan, 1974). It seemingly measures active comprehension as students must attend to the meaning of the passage while they read,

clarifying mistakes by choosing the correct word from each of the choices. As students encounter each blank, they are able to both reread prior text and scan ahead to discern the meaning and select the correct word. Thus theoretically, MZ is a quantifiable observation of vocabulary knowledge and syntactic understanding as well as students' ability to employ active reading strategies for meaning monitoring while they read (Howell & Nolet, 2000).

MZ is a commonly used assessment strategy because it is easy to create and administer frequently for the formative assessment of reading proficiency. Evidence suggests MZ is a reliable tool for assessing intra-individual differences regarding student reading achievement, and it is sufficiently sensitive in measuring growth over repeated administrations (Shin, Deno & Espin, 2000). This same study also found MZ predictive of student achievement differences. Interestingly, research indicates MZ reflects student growth where ORF begins to wane in sensitivity (Jenkins & Jewell, 1993). In fact they found MZ to be more sensitive to individual differences at upper grades than they did at lower grades.

More recently, Retell fluency (RTF) has gained attention since it was added to the sixth edition of the DIBELS standard battery in 2002. The authors described its usefulness as adding face validity to the construct of oral reading fluency (Good & Kaminski, 2002) and suggest it provides a good indication of a student's overall reading proficiency and reading comprehension abilities. As designed, the DIBELS retell is a fluency measure. After the student reads a passage aloud for one minute to get a ORF score indicated by words read correctly, the student is asked to retell everything he or she

remembers from the passage. For one minute the students retells all he or she can recall while the test administrator counts the words retold omitting repeats and nonwords.

There is a dearth of published information supporting RTF reliability and valid measure of reading comprehension. In initial studies inter-rater reliability was observed at about .96, alternate form reliability and test-retest reliability at less at .47 to .70 and .58 to .78, respectively (Manchester, Marcotte & Matthews, 2004). Little is known about the validity of this assessment, nevertheless retell fluency is worth exploration as a simple and efficient tool that could be given in concert with a simple read aloud reading fluency task. Thus if retell does add to our understanding of reading proficiency, this measure is convenient to add to a battery of formative assessments.

Fuchs, Fuchs and Maxwell (1988) found written retell (WRT) to be a more acceptable procedure for assessing reading comprehension than oral retell. These researchers observed, "written rather than oral production method produced stronger correlations with the criterion reading comprehension test (p. 26)," the Reading Comprehension subtest of the Stanford Achievement test, 7th edition (Gardner, Rudman, Karlsen & Merwin, 1982). Compared with the correlation of .59 total words orally retold and .64 content words retold, content words written correlated at .82. While this research indicates WRT may be a valid formative measure of reading comprehension, it is not widely used as a formative assessment tool for reading comprehension. Fuchs, Fuchs & Hamlett (1989) found WRT to have instructional utility through their study that demonstrated when teachers used student recalls as an ongoing measurement system, they utilized the results to set ambitious goals and modify their instructional activities in accordance to their students needs.

While a WRT procedure could be an efficient assessment to employ regularly in classrooms, currently it is not available in a standardized format and many scoring procedures are cumbersome and time consuming. If WRT is found to provide teachers useful information above and beyond the tools more readily available to them, it is important to standardize this procedure and thus utilize it as a systematic assessment tool

In this study, the WRT procedure will be defined as a five-minute passage read followed by a five-minute written recall, as described in the Fuchs, et al. (1988) study. To further standardize the procedures, students will be prompted to continue writing all they can remember at each one-minute mark. Content words counted for the retell score are defined as all distinct proper and common nouns, verbs, adverbs, and adjectives the same or synonymous with those in the passage. Misspellings will be counted as long as 50% of the letter sequences are spelled correctly (White & Haring, 1980).

Another measure, the Sentence Verification Technique (SVT) is less widely used as a formative assessment measure, but may be the best way to directly observe constructing meaning from a linguistic string within a text by it's interaction with what the reader already knows (Royer & Cunningham, 1981). Materials for this testing procedure can also be taken directly from the curriculum so as to be a direct measure of students' comprehension of classroom materials. In this assessment, students are provided three to four short 12-sentence passages followed by a set of test sentences. There are four types of test sentences, two that mean the same thing as the sentences in the passage and two that have different meanings. Same meaning sentences include original sentences that are taken directly from the passage and paraphrase sentences where as many words as possible are changed while maintaining the passage meaning.

The two sentence types with different meanings include a meaning change sentence where as few words as possible are changed so as to change the sentence meaning, and a distractor sentence that is similar only in sentence structure, but not words nor meaning. After reading the passage, the student must decide whether each sentence means the same thing as the passage. They must indicate a "yes" or "no" for each test sentence. Whereas MZ is criticized for measuring only sentence understanding, SVT has been found a good measure of full passage comprehension (Royer, unpublished manuscript).

Purpose of this Study

In addition to reading fluency, this study focuses on the incremental and predictive validity of these four formative measures of comprehension - MZ, RTF, WRT and SVT. These measures have been identified for their promise as efficient and reliable measures of reading comprehension. They have been selected for investigation in this study because of their potential utility in four important areas as assessments of reading comprehension. They each could be used formatively, given as frequently as necessary to inform instructional decisions. They each show promise in quantifying possible indicators of reading comprehension proficiency, and they are either already commonly used or could easily be employed as assessment strategies used by classroom teachers. Finally, they show promise for meeting the specifications for CBM as indicated by Deno (1985), whereas they are inexpensive to produce and efficient to administer and can be easily understood by school-based consumers. They measure indicators of academic health based on production-type responses, are sensitive to growth over time and as a result of instruction and are relevant to instructional content.

Little is known about the construct utility of these assessments and how they add to our understanding of the development of reading proficiency. Initially, it is necessary to understand how they each contribute as indicators of overall reading proficiency and comprehension development. Once such foundational information is understood, further investigation regarding sensitivity to instructional interventions and the utility of each as formative assessment measure can be explored.

The purpose of this research is to investigate the incremental utility of formative assessment measures of reading comprehension. It is hypothesized that, due to the many subskills subsumed within the process of reading, formative assessment measures of reading comprehension will contribute more to our knowledge of reading skills than simply oral reading fluency can provide. Furthermore, it is hypothesized these measures can be modeled in a meaningful way, and this meaningful model will predict performance on overall measures of academic success.

Research Questions and Hypotheses

Question 1: Accounting for variability in the construct of reading comprehension.

Do formative measures of reading comprehension account for the variability of reading comprehension beyond the variance accounted for by oral reading fluency? It is hypothesized that existing formative assessment measures of reading comprehension, the Maze procedure, Retell fluency, written retell, and SVT will account for significant variability in overall comprehension ability beyond the variance accounted for by oral reading fluency.

Question 2: Predicting academic performance

Can this model of reading comprehension measures reliably predict academic achievement as indicated by pass or failure on the statewide assessment beyond the predictions made of oral reading fluency? It is hypothesized that the identified model of reading comprehension will be a better predictor of success on the statewide assessment than the use of oral reading fluency as the sole indicator.

CHAPTER 2

REVIEW OF THE LITERATURE: THE RESEARCH ON READING COMPREHENSION INSTRUCTION AND ASSESSMENT

The Current State of Literacy in America

As a result of the continuous evolution of evidence in best practices for teaching students to read, scientists estimate we now know enough about reading instruction so as to ensure at least 95 percent of the children in the United States are taught to read at a level of proficiency so as to enjoy and engage in independent, age-appropriate reading activities (Moats, 1999). Despite what we now know about reading instruction, there has been little evidence of improved literacy outcomes in American students in the last thirty years (NAEP, 2005). According to the American Federation of Teachers, about 20 percent of elementary school students have significant problems learning to read and an additional 20 percent do not read well enough to enjoy independent reading (Moats, 1999). Even more striking is the growing achievement gap between students with minority racial status and those in the majority in this country. In fact, the percentage of minority students unable to read is greater today than it was at the beginning of the 20th century. With children from minority backgrounds and impoverished homes having a rate of reading failure between 60-70% despite great efforts from Title I and other expensive and extensive initiatives (Sweet, 2004). The implications of illiteracy have weighty implications associated with poverty. Adults who can not read are more likely to live below the poverty lines due to unemployment and underemployment. In fact, with 25 percent of adults in this country lacking the basic literacy skills required to be successful in a typical job (Moats, 1999), the implications of illiteracy have been designated a public health concern.

These statistics stand in stark contrast to what we now know about reading instruction and the prevention of reading failure. With good reading instruction that targets direct teaching of phonemic awareness, systematic and explicit phonics instruction, vocabulary instruction that integrates a variety of methods to make meaning connections, and direct instruction in the comprehension strategies that include clarification, question generation and summarization, as well as daily exposure to rich text and writing activities, we have the knowledge and ability to nearly ameliorate reading failure in the United States. (NRP, 2000) "In the field of reading, we now have definitive answers to some age-old problems that, if applied, could transform classroom education and virtually eliminate illiteracy in America (Sweet, 2004, p.14)."

Literacy Instruction and the Federal Government

In the midst of the educational debates and cross discipline "Reading Wars" (Chall, 1967), the irony between what we know is possible and the present state of literacy instruction in American education was brought to the attention of political leaders, legislators and subsequently, to the American public. In his 1996 State of the Union Address, President Clinton outlined the stark consequences of the American reading deficit, bringing reading instruction to the forefront of the federal government and its policy makers (The White House, 1996).

As a result, the House of Representatives enacted the 1998 Reading Excellence Act, which had two important consequences. First, the Act allocated more than \$200 million annually to help state departments of education acquire instructional materials and assessment tools and provide teacher training with the most current research on effective reading instruction. Secondly, Congress applied the term "scientifically based

reading instruction” to the parameters of such materials and professional development, a term that would take on a life of its own as the monies and intentions of the Act rolled out to school districts and educators. For the first time in history, the federal government legislated the standards for literacy instruction by defining the term “scientifically based reading instruction” to mean

(A) the application of rigorous, systematic, and objective procedures to obtain valid knowledge relevant to reading development, reading instruction, and reading difficulties; and

(B) shall include research that –

- i. employs systematic empirical methods that draw on observation or experiment;
- ii. involves rigorous data analyses that are adequate to test the stated hypotheses and justify the general conclusions drawn; and
- iii. relies on measurements and observational methods that provide valid data across evaluators and observers and across multiple measures and observations; and
- iv. has been accepted by a peer-reviewed journal or approved by a panel of independent experts through a comparably rigorous, objective, and scientific review. (“Reading and Literacy Grants to State Educational Agencies,” Title II (C) Sec. 2252 (5) [20 U.S.C. §6661a])

In a radio address to the nation in 2001, President George W. Bush placed early literacy instruction at the forefront of his administration’s priorities and set the goal that “no child shall be left behind because he or she can not read.” In his address he delineated the three critical methods to meet the challenge. First, he called upon the use of valid diagnostic assessments to identify students who fail to develop strong literacy skills as early as possible. Second, he proposed the use of early interventions to correct

reading problems before they are too difficult to remedy. Lastly, he stressed the use of phonics as having a central role in early literacy instruction. Four months after his address to the nation, President Bush signed the No Child Left Behind Act of 2001 (NCLB) into law. In NCLB there were 110 references to “scientifically based reading instruction,” and for the first time, essential components of literacy instruction based on three decades of research were written into federal law.

The primary goal of NCLB is to close the literacy gap between all students and all schools regardless of the socio-economic areas from where schools draw their populations. There is immense socio-political pressure for educators to eliminate the achievement gap that exists between minority and majority students and ensure every child is reading on grade level by the time they leave the third grade. Educators and district administrators are being held accountable for the educational outcomes of their students more than ever before by both legislators and by the public. The primary vehicle of defense to show a school district is doing everything in its power to educate every student who crosses its threshold has been the adoption of science-based instructional practices.

“Scientifically-based Educational Research” and Practice

“Science-based” has become the new catchphrase in the field of education as its meaning becomes distorted by retailers and marketers of educational materials and misunderstood by consumers, parents as well as educators and administrators. A walk through a teacher supply store reveals “science-based” scribed on the majority of the instructional materials. It has become difficult for the untrained eye to distinguish between instruction based on rigorous scientific evidence and that which is not,

complicating the difficult task of providing effective instruction and raising the achievement levels of all students.

Historically, the field of education has been flooded with innovations that did not necessarily improve academic outcomes. Movements in the field of education tend to flit from one romantic notion about learning and child development to another while educational researchers try to incorporate what is known about effective instructional practices. When large-scale research efforts irrefutably reveal the greatest gains in academic outcomes were the result of direct instruction practices over practices based on a more natural view of learning, the results were largely criticized as too statistical (Grossen, 1995-96). Instead of embracing the exciting information regarding best practices for teaching all students, critics called for a case study analysis to describe whether less effective natural-learning based programs had a positive impact on teachers without regard for academic outcomes (Glass & Camilli, 1981). As a result the conclusive evidence about best instructional practices was poorly dispensed to practitioners and when done so, the effects of the lesser practices largely exaggerated (Watkins, 1995-96).

Often, it is prudent to rely on experts in a professional field to disseminate research findings regarding best practice. It is their job to be objective and base their decisions on research and not on ideology. We rely on experts to publish the results of their findings only after colleagues of the respective fields formally and critically review those results. We expect our experts to put their ideas out in the public so as to be tested by others and make public when they were wrong so as to move forward with confidence. Only when false theories are publicly debunked, valid theories replicated and all

information widely disseminated will a field mature into a profession where its practitioners base their decisions on science and research. In 2000, Doug Carnine warned, "Until education becomes the kind of profession that reveres evidence, we should not be surprised to find its experts dispensing unproven methods, endlessly flitting from one fad to another. The greatest victims of these fads are the very students who are most at risk (p.3)."

While it is arguable whether the field of education is a mature profession that relies on convergent evidence, there still exists the foreboding task of sifting through the multitude of information regarding effective instructional practices so as to provide every student optimal educational opportunities. Educators and school administrators can be empowered to provide their students effective instructional practices by evaluating what they do in three ways: (a) through formal testing and student evaluations procedures, (b) through published evidence that the instructional approaches have led to student achievement, and (c) when direct evidence is not available, using reason to rely on practices that converge with the science literature. (Stanovich & Stanovich, 2003).

Convergent Evidence about Literacy Instruction

Thirty years of reading research has produced indisputable evidence regarding the most effective practices for beginning reading instruction. Consistently, convergent evidence has shown direct, explicit and systematic phonics instruction for struggling learners is significantly more effective than unsystematic phonics instruction or no phonics instruction (Ehri, 2004). Evidence also suggests phonemic awareness, the ability to separate sounds from whole words, is highly predictive of early reading achievement (Share, Jorm, Maclean, & Matthews, 1984). Most importantly, when children are unable

to distinguish each sound in spoken words, phonemic awareness can be taught such training has resulted in significant achievement in reading acquisition (Ball & Blachman, 1991). Research also reveals that fluency is a critical feature of successful reading, indicative of facile word recognition skills and contextual understanding (Fuchs, Fuchs, Hosp & Jenkins, 2001).

Not only have education researchers been able to identify phonemic awareness, phonics and fluency as critical features of effective early reading instruction, they have also conducted innumerable studies to address the most effective instructional practices for teaching these three components to all learners. This research has provided educators a scope and sequence of important instructional elements for teaching phonemic awareness, phonics and for building fluency for each subskill and in reading in general. Furthermore, educational research exploring how best to teach these elements reveals that direct and systematic instruction accelerates the learning process for all students, particularly those students at greatest risk for school failure.

Direct instruction (DI) is based on theory that teacher-directed lessons presented clearly and systematically so as to preemptively reduce any confusion for students designed with scaffolds to facilitate generalization is the best instructional approach to accelerate the learning process (Carnine, Silbert, Kame'enui & Tarver, 2004). Most DI lessons are designed to present new information in a three step process where the teacher models the new information, leads the students in exploring and practicing the new information and then tests the students for their independent understanding. Often DI lessons are scripted, always they are designed with a specific sequences of skills with each new lesson built upon the previous lesson, thus focusing on mastery of basic skill

knowledge and application before moving on to higher-order skills, such as creative problem-solving.

Importantly, DI is based on three underlying assumptions (Gersten, Woodward & Darch, 1986). First, DI lessons are highly structured regarding scope, sequence and content. It is assumed learning basic foundational skills and then their application to higher order skills is essential to complex thinking. The second and third assumptions explain why DI strategies have become important elements of school reform and remedial instructional programs in the age of the No Child Left Behind Act. DI assumes all children can be taught and students who are disadvantaged must be taught at an accelerated rate compared to that of their typical peers so as to close the achievement gap that would continue to expand if not remedied.

The State of the Research Regarding Reading Comprehension

Whereas evidence regarding the important instructional elements and best ways for teaching students beginning reading skills, particularly lexical access to print, has been replicated by independent researchers creating an accumulation of convergent scientific evidence, we have less conclusive information regarding instructional approaches to teach children how to extract meaning from text. Touted the fundamental goal of reading, reading comprehension is undoubtedly an essential feature of good reading instruction, in addition to phonemic awareness, phonics, fluency and vocabulary development. To date, two large-scale research projects have conducted meta-analyses on the research base to uncover empirical evidence to direct our efforts in reading comprehension instruction (NRP, 2000; RAND, 2001). Despite these massive projects, the state of the research and our foundational knowledge regarding the important

elements involved in higher order reading skills necessary for extracting meaning from text are much less understood.

Not much has changed since Durkin responded to a 1976 Request for Proposals (RFP) from the National Institute of Education. At the time NIE was looking for a center to explore best practices in reading comprehension. Cited in the background and purpose of Durkin's 1976 study was the explanation of the NIE for supporting reading comprehension research

A considerable, though not entirely adequate body of facts has been assembled about decoding but much less is known about the process of understanding written text. Researchers and practitioners accordingly, have strongly urged the NIE to focus its attention and that of the field on the problems of reading comprehension (p. 2)

Durkin decided before a scientific evaluation of current instructional practices of reading comprehension could commence, it was necessary to conduct preliminary investigation to explore what, if any, reading comprehension instruction was currently taking place in American classrooms. In 1979, she published the results of her observational study examining the amount of reading comprehension instruction conducted in a sample of American classrooms at the time. Durkin found only 2% of instructional time was spent on the direct teaching of reading comprehension strategies. Unfortunately, twenty years later the situation has not changed where even award-winning teachers were observed asking more comprehension monitoring questions rather than specifically teaching comprehension strategies. (Wendler, Samuels & Moore, 1989). For the direct instruction of reading comprehension to become a more meaningful part of American classrooms, it is important to provide teachers with specific strategies and instructional approaches that result in meaning gains in reading comprehension.

Reading researchers continue to seek out the preeminent instructional strategies for teaching higher order reading skills, such as the factual and procedural knowledge necessary for developing strong vocabularies and robust comprehension skills. In addition to reading comprehension instruction evaluated within the NRP, the RAND Reading Study Group (RRSG) also studied our current understanding of the construct. The study group laid out a research agenda based on what they described as a “fairly well articulated knowledge base” from which they delineated ten things we know about reading comprehension and nine important questions that still need to be answered.

Of the ten items the RRSG identified, for only five do we truly understand their impact on instruction. We know that instruction designed to develop fluent word recognition has moderate gains on reading comprehension. It has also been found that instruction directly teaching students cognitive strategies may enhance their ability to monitor their understanding about what they read (Gersten, Fuchs, Williams & Baker, 2001; NRP, 2000; Rosenshine & Meister, 1994), and best results for struggling learners occur when this instruction is direct and explicit (Wong & Wilson, 1984). It is also clear when students do not understand text structures and grammatics, they are ill equipped to organize the information they read, and explicit instruction in text structures give students a framework for arranging new information for meaning. (Meyer, Brandy, & Bluth, 1980). Finally, giving students choices about what they will read and interesting texts to choose will encourage their motivation and task persistence in the task of reading (Reeve, Bolt, & Cai, 1999).

The remaining points delineated in the RRSG research agenda regarding what we know from reading comprehension research have implications for future research.

Specifically, we know that vocabulary plays an extremely important and equally complex role in comprehension. We also know there are numerous hypotheses regarding best ways for teaching reading comprehension. These two points hardly contribute to our understanding of comprehension instruction, rather to our understanding of all we have yet to discover.

The RRSg then listed all we have yet to discover about reading comprehension instruction. We need to investigate, whether allocating more time, using current curricula and comprehension strategies could provide struggling comprehenders sufficient gains, and what strategies, information, and instructional approaches should best be used. In summary, the RRSg's questions to promote a future research agenda in the field of reading comprehension implies we know little about what to teach and how to teach it to generate meaningful gains in comprehension instruction.

Alas, more than 25 years after the NIE called for research on the construct of reading comprehension, the RRSg described the rationale for the next steps in their research agenda that nearly parallels that cited two and a half decades before:

Recent research on reading instruction has led to significant improvements in the knowledge base for teaching primary-grade readers...Nevertheless, evidence-based improvements in the teaching practices of reading comprehension are sorely needed...[and] should be the primary motivating factor in any future literacy research agenda (p.xi)

Evidence in Reading Comprehension Instruction

With respect to reading comprehension instruction there are three important distinctions to consider: (1) what the student brings to the task of reading, (2) strategies good readers employ to facilitate text understanding, and (3) the best instructional

methods for teaching effective cognitive strategies. Good reading comprehension instruction fortifies the foundational enabling skills students bring to the task of reading so they find success during reading and increase their motivation to read more. Good reading comprehension instruction directly addresses the cognitive activities good readers bring to the task of reading. Most reading comprehension research has investigated the many possible cognitive activities to determine which are the most effective to construct meaning. Finally, cognitive strategies students might learn is often confused for instructional approaches teachers might use to teach the cognitive strategies (Dowhower, 1999). Good reading instruction entails presentation of strategies addressing every phase of the learning hierarchy (Haring, Lovitt, Eaton & Hansen, 1978) for every enabling skill and every higher order cognitive strategy (Carnine, Slibert, Kame'enui & Traver, 2004). How cognitive strategies are taught is just as important as what cognitive strategies are taught.

The learning hierarchy model suggests students acquire new skills in a predictable sequence (Haring, Lovitt, Eaton & Hansen, 1978). First they acquire basic understanding of the new skill, attempting to apply it as they learned, but often making mistakes in their attempts. Once students acquire a skill they tend apply their new knowledge methodically and very accurately. This stage of accuracy requires great cognitive attention to the task. Once students obtain accuracy, they become fluent in a skill, applying it both with accuracy and automaticity through. Once students reach the stage of fluency, the new skill requires little cognitive attention. Finally, students begin to apply their new knowledge to new contexts and settings.

Specifically, instruction to address both the enabling skills that allow for effective reading and the higher order skills to elicit effective meaning construction should be directly taught and meticulously modeled in a systematic way so students can accurately acquire the new behaviors as they are taught. Instruction should then provide opportunities for guided practice for students to grasp the skills with accuracy and consequently fluently. Once students become automatic in their application of the new skills they have learned, effective instructional strategies should allow independent practice in a variety of contexts so that students can generalize their new skills to new reading experiences.

Enabling Factors

Children bring with them enabling skills to access the print off the page through their agility in word recognition and then their ability to access individual word meanings through the complexity of their vocabulary knowledge. Both facile word recognition skills and access to word meanings play important roles in the task of text comprehension (Howell & Nolet, 2000). Children also approach text reading with a prior knowledge base that may facilitate understanding but may also hinder it if the text conflicts with information the reader already knows. Not only are there differences between individuals in their knowledge base of text content, there are also differences within individuals as they move from one text to another (RAND, 2002). Finally, students bring to every task of reading different levels of motivation and persistence. Children with interest in the text often are highly motivated to read even when the task may be difficult. Background knowledge increases with reading and reciprocally facilitates understanding. The more students read, the more they know, the more enjoyable the task of reading and then the

more the child is likely to read. Inversely, the more a student struggles to access print, the less vocabulary and background information they acquire, the less they know that facilitates understanding and the less motivated they are for reading (Stanovich, 1986). Thus task persistence and motivation are reciprocally necessary for successful reading and reinforced through successful reading.

Instruction that prepares students for comprehension of text includes providing them with facile word recognition strategies, robust vocabulary development and a depth of background knowledge. Equally important, instruction to reinforce motivation and task persistence should be designed to ensure students consistently have successful interactions with text as they are developing readers. Repeated failures with reading will decrease motivation to read and to endure through difficult reading tasks. Most children with reading problems have limited persistence to maintain attention when reading (McKinney, Osborne, & Schulte, 1993).

Cognitive Strategies and Teaching Approaches

Most research regarding effective comprehension instruction has focused on cognitive strategies that best facilitate understanding through shaping students' active roles in constructing meaning during reading (Lysynchuk, Pressley, d'Ailly, Smith & Cake, 1989). The goal of cognitive strategy instruction is to provide students with a specific set of organized activities that facilitate understanding and aid problem-solving on intellectual tasks (Torgesen, 1982). It is theorized that during reading, people engage in a variety of constructive processes that begin with the lexical input and only evolve into meaning through active and purposeful employment of these cognitive processes (Kinstch, 1988). Thus, meaning is constructed as a result of an interchange between what

the author intended to express in the content of the text in combination with the prior knowledge the reader brings to the task of understanding.

Implied in the theory of active meaning construction is the idea that poor comprehenders do not adequately utilize effective cognitive strategies to construct meaning (Dowhower, 1999). Historically, the goal of comprehension instruction has been to provide readers with a set of cognitive strategies that induce purposeful attention to the words they read and problem-solving strategies when there is an obstacle to understanding. The intent of such instruction is to help make students aware of their own thinking through teacher modeling which would then gradually be released to the students as they internalize the use of the strategies (Pressley, Johnson, Symons, McGoldrick, 1989).

The last three decades of scientific exploration around effective comprehension instructional practices has filtered a broad array of teaching recommendations into practice through scientific journals and expert recommendations. Often evidence revealed one effective strategy or another, but there has been limited convergent evidence and some serious methodological flaws in the research (Lysynchuk, et. al., 1989). Various large-scale methodological evaluations and meta-analyses helped practitioners make sense of the sparse evidence supporting broad recommendations (Gersten, et. al., 2001; NRP, 2000; Lysynchuk, et. al., 1989).

With relatively new federal laws supporting the use of only scientifically based practices for teaching literacy in America and the legislated guidelines for evaluating educational science, the National Reading Panel investigated the breadth of scientific literature around comprehension strategy instruction in an attempt to sift through the

published recommendations and provide educators with a blueprint of the most effective approaches to teach reading comprehension. The study group cast embarked on a comprehensive search and found 481 studies on reading comprehension instruction over the last three decades. Limiting their evaluation to studies that did not include content areas such as science and mathematics, only studies published in peer-referenced, scientific journals that included an experimental group with a control group for comparison, the NRP sifted through the studies and ended up with 205 studies that reported effective strategies for teaching comprehension.

Their investigation revealed 16 different categories of instruction that had been published in scientific journals claiming to have positive effects on reading comprehension. Of those 16 strategies, the NRP concluded only seven categories of instruction “appear to have a firm scientific basis for concluding that they improve comprehension instruction in normal readers (NRP, 2000, p. 4-42).” Of the seven categories that were identified, five categories specifically address cognitive strategies that appear to result in meaningful gains in comprehension. The NRP found that (a) teaching students to monitoring their own understanding while they are reading, (b) to generate questions and (c) answer questions during and after reading, (d) to summarize what they read, and (e) to make use of multiple cognitive strategies while they reading lead to positive gains in comprehension. The remaining two of the seven categories describe methods of teaching reading comprehension that seem to show promise, specifically the use of graphic and semantic organizers and the use of reciprocal teaching or collaborative learning strategies.

Comprehension monitoring strategies. In its broadest definition, comprehension monitoring is the awareness of one's own understanding and the cognitive activities one might engage when actively searching for meaning. Most comprehension strategies can broadly fit under this definition because the goal of all strategies is to identify and construct meaning from new information. More specifically, however, when examining each specific strategy, the goal of pure comprehension monitoring strategy instruction is to help students develop metacognitive awareness when they encounter an obstruction to effective understanding and provide students the means to fix the misunderstanding. The purpose of comprehension monitoring instruction is to teach students to have autonomous cognitive control and self-regulation while they are reading in which they are able to adjust their own reading processes in accordance to their ongoing understanding. Steps of such instruction often include teaching students to become aware of what they understand, to identify what they do not understand, and to repair confusions so as to create meaning (Baumann, Jones, & Seifert-Kessell, 1993 & Paris, Lipson & Wixon, 1983).

Comprehension monitoring strategy instruction is rooted in early research exploring variability in children with learning disabilities. Historically, the research sought a crystallized deficit in cognitive processing to account for differences in children with learning disabilities compared to more typically performing students. In 1977, Torgesen published a relatively novel perspective of the learning disabled child, asserting learning disabled children did not necessarily have concrete cognitive deficits, rather they were more likely to be "inactive learners," passively perceiving new information without the purposeful integration of the information into their knowledge base as do

more typically performing students. He and others (Tarver, Hallahan, Kauffman, & Ball, 1976 & Wong, 1980) demonstrated struggling learners did not employ goal-directed strategies to aid them in tasks of perception, attention and memory. Importantly, much of the same research demonstrated when students are taught specific strategies for active learning, they tend to use them. These results led to inferences that students with learning disabilities do not have crystallized cognitive deficits, but need more instructional scaffolding to employ active thinking strategies. Eventually, this line of research led to the idea all students can benefit from instruction in active awareness of their own thinking processes, specifically when reading.

Comprehension monitoring is a strategy students need to employ while they read as they actively interact with the text. The development of an awareness of one's own thoughts tends to evolve in skilled readers who are more likely than poor readers to employ flexible, and purposeful thinking to self-monitor as they read. Self-awareness has been observed as an ability that develops over time in typical learners, however, it has been shown that direct instruction of self-monitoring strategies can benefit all learners (Paris & Jacobs, 1984).

Very often comprehension monitoring is taught by teachers questioning and commenting during reading with the goal to have children attend to the text content. Similar to the observations Durkin (1979) made more than 25 years ago, it has still been observed that teachers who believe they are teaching comprehension strategies are merely assessing their students' comprehension through this method (Wendler et al, 1989). The goal is to provide the right amount of scaffolding so as to shift the self-monitoring strategies to the students.

Teaching students to think-aloud while they read has been found to help students stop when they fail to understand what they are reading and has been able to provide students with various strategies to repair the misunderstanding. In one study, students were taught when to ask themselves questions, how to stop and self-question when they come across inferences in the text, to restate what they read in their own words so as to facilitate understanding, and rereading strategies. Children who received the think-aloud instruction were more likely to verbalize what they were thinking while they read when probed by an interviewer than were the children who did not receive the think-aloud training (Baumann, et al., 1993).

Building on the idea that teaching children to monitor their own comprehension will result in active and purposeful readers, more instructional strategies have been filtered into classrooms with the goal of scaffolding independent metacognition and problem-solving during reading. The "Click and Clunk" is one of those commonly used approaches. This strategy teaches students to stop after each sentence and ask themselves, "do I understand what I just read?" If the student understands, then they say "Click" and move on. If they do not understand, they say "clunk" and refer to the fix-up strategies worksheet that provides them options to fix word confusions and sentence confusions, as well as checking for paragraph and page meanings (Klinger, Vaughn, Dimino, Schumm & Bryant, 2001).

The majority of studies examined by the NRP investigated whether specific instructional strategies could help children monitor their own understanding and employ fix-up strategies independently. Research suggests direct instruction focused on teaching children to become aware of their own thoughts and misunderstandings while they read

seems to improve students' skills that specifically relate to self-monitoring, such as error detection, recall, and answering specific questions about what they read. There is less evidence to suggest that students transfer their learning to new tasks, especially for younger students (NRP, 2000). It appears direct instruction of comprehension monitoring strategies can result in students being better aware of their understanding and lack of understanding while reading. Such instruction can also effectively provide students with strategies to repair misunderstandings. The NRP concludes, "[Comprehension monitoring] may be a useful addition to a program of instruction that employs flexibility and the teaching of multiple comprehension strategies (NRP, 2000, p. 4-44)."

Summarization. Teaching students to summarize what they read has been a popular strategy for engaging students in the meaning of text through instruction around main ideas and themes. A summary is simply a brief account of the most relevant elements of a longer text. The goal of summarization instruction is to teach children to detect the most important features of text by sifting through trivial details and redundancies so as to identify the central ideas.

Instruction around summarization arises from a solid cognitive theoretical base that purports a personal summary of a text naturally occurs as a result of understanding (Kintsch & van Dijk, 1978). These theorists defined understanding as a three part system. First readers perceive input from the details and relationships between the details in the text. These text details were called microstructures. Readers then apply a set of rules to create meaning from the microstructures. The rules are weighted specifically to each individual reader according to what the reader already knows, the purpose for reading the text and expectations they bring to the task of reading. The elements of this second set of

operators allow the reader to get the gist of the text, or the macrostructure of the text.

The final process includes steps in which the reader essentially generates a new cognitive representation of the original text from what they remember of the microstructures and the gist. This new representation can be viewed as a personal summary of the original text. Based on this theory, instruction focused on teaching students to summarize text would consequently be directly teaching students to understand.

An analysis of studies that examined summarization as an effective strategy to aid comprehension revealed direct and explicit instruction in a specific set of steps, allowing for the students to master each step resulted in the best outcomes, including better summarizations, memory of text details and application of strategies to novel situations (Hare & Borchardt, 1984). Brown and Day (1983) designed rules for training students in effective summarization that correspond with the macrorules Kintsch and van Dijk (1978) theorized happens automatically for mature readers. The six rules for generating text summaries include, 1) deletion of trivial information, 2) deletion of redundant information, 3) the inclusion of superordinate terms to describe for lists or series within the text, 4) the inclusion of superordinate terms to describe the actions and events within the text, 5) selection of a topic sentence from the text, or 6) formulation of a topic sentence if one is not explicitly stated. Studies that have used these steps or steps modified from these have shown positive effects when they used direct instructional strategies to train their students (Bean & Steenwyk, 1984; Hare & Borchardt, 1984; Rinehardt, Stahl, & Erickson, 1986).

Consistently results of these studies show students who are systematically trained to summarize what they read are able to produce better written summaries than those who

are not trained. Predictably, some evidence shows summarization instruction improves students' retention of the big ideas of text, yet not details of what they read and often result in improved study skills (Brown, Day, & Jones, 1983; Rinehardt, et al., 1986). In their analysis of summarization strategy instruction, Pressley et al. state, "Finally, nothing is known at this point about the long term benefits of summarization training.

Nonetheless, the evidence to date in favor of this strategy as a facilitator of comprehension and memory is so striking that we recommend the procedure without hesitation (p. 9)." Less vehemently, the NRP also recommends summarization as an important part of reading comprehension strategy instruction.

Question generation and question answering. Most often, the purpose of question and answering instruction is to encourage students to attend to the facts and authors intent within the context of a text and consequently encourage active student engagement during text reading. Most commonly, teachers pose questions to their students before reading to help focus the students' attention to the main ideas that will be presented in the text through the text structure, vocabulary or story's theme. In some instances, teachers have employed the questioning strategy to encourage their students to attend to important features of the text during reading by imbedding questions into text.

Embedding questions into the text is most commonly used during text read-alouds, and is so popular as to be an important feature within the teacher's editions of most commonly published literature curricula. Good text talk requires teachers to ask more than factual questions about what information is in the text, but also inferential questions that require students to recall things they already know and integrate them with information from the text. Often teachers will frame their questions around a child's first

response and encourage them to expand upon what they initially knew. (Beck & McKeown, 2001).

With the premise that students with learning disabilities tend to utilize less active and purposeful strategies to construct meaning while reading, Wong (1980) devised two studies to test the hypotheses that 1) students with learning disabilities do not automatically employ processing strategies to aid in text understanding and 2) with instruction these students can and will use such strategies thus implying they have the cognitive ability to do so, and merely require more specific instruction in attending to text. Wong (1980) found that children with learning disabilities did more poorly than typical readers when the consequences of the sentences were implicit, yet performed equally well when the consequences were explicitly stated. She also found that students in grade two performed significantly worse on the same task than did students in the sixth grade. The first results confirmed that students with learning disabilities did not actively construct meaning whereas their typical peers had. The second result reflected the developmental nature of meaning construction.

In the second part of her study, Wong provided struggling learners with questions and prompts to recall the inferred consequences. With the prompts, the students performed more similarly to the better readers, thus demonstrating that learning disabled children so have the cognitive abilities to infer meaning from text, and using questioning strategies can help them do so.

While the NRP found positive effects with question-answering strategies in facilitating comprehension, the best results were found when teachers provided corrective and supplemental feedback. However, the research shows gains made as a result of

question and answering strategy instruction were limited to the specific texts where the questions were posed and did not generalize to other situations. In contrast, teaching students to generate their own questions while they read showed the greatest positive effect on comprehension compared to all the strategies deemed effective by the NRP.

Typical instruction in question generation entails teaching students to ask their own factual and inferential questions while they read and then encouraging them to seek answers to their own questions. Students are often taught to generate questions as part of their clarification strategies, designing questions about things they do not understand within the text that require further investigation. More commonly students generate questions factual questions using common who, what, where and how questions starters and why as the inferential question starter. These approaches to instruction of question generation are watered down from their original intent.

The original idea was to teach students to generate their own questions before and after reading. This strategy instruction is based on the theory that reading comprehension is an interactive processes where meaning is constructed as an interaction between resources the reader brings to the activity and characteristics of the text (Kintsch & van Dijk, 1978). The goal of question generation strategy instruction is to teach students elements of text and the use of contextually relevant details to map onto questions that encourage students to be attuned to important story elements, develop schema based on larger contexts and understand what details are unique to the text. Theoretically, the result is to produce more reader-based understanding of the answers than answers that rely mainly on the text itself (Singer, 1978).

Singer and Donlan (1982) taught students specific story elements, including character, goal, obstacles outcome and theme, along with general corresponding questions. During each instructional session students in the experimental condition were taught a new question framework, asked to write questions to a story they would hear, then prompted to write more questions at the end of the story. Students in the control condition were asked teacher generated questions. Each group took a 10-question test after each story. After the first two trials, there was an increase in performance of students in the experimental condition and a significant effect over all the testing sessions. The researchers concluded that their training, which combined teaching important story elements in addition to question generation, resulted in students using more reader-based resources than text based resources in processing text.

Of the 27 studies evaluated by the NRP that explored question generation strategy instruction, 16 resulted in significantly better performance within the experiment conditions on experimenter designed tests with effect sizes ranging from .85 to .95. Positive gains were also observed on standardized comprehension assessments, although the effect size of .36 was not statistically significant raising some doubt to the generalizability of this approach. Of all the strategies deemed effective by the NRP, teaching students to generate questions while they read showed the strongest empirical support.

Multiple Strategy Instruction and Reciprocal Teaching. Being a purposeful and strategic reader requires actively seeking out meaning by any means available. Thus, comprehending text requires much more than knowing the individual strategies. Good readers coordinate many strategies and shift from one to another as it is appropriate to do

so according to the specific problem encountered during the reading process. Good readers will constantly adjust the strategies they use until the obstacle to comprehension is resolved and meaning is constructed. The NRP includes multiple strategy instruction as one of their recommendations for effective comprehension instruction. The goal of multiple strategy instruction is to provide students a cache of strategies to draw upon to accommodate the various problem-solving situations any reader might encounter. There is promising evidence that providing students with a variety of strategies to choose from is more likely to lead to application of their new knowledge to novel reading tasks than instruction in just one strategy.

Schumaker, Deshler, Alley, Warner and Denton (1982) taught students with learning disabilities a complex series of strategies to employ before after and during reading. Students were taught to use the strategies through three passes of each text unit. In the first pass of the text, students were taught specific ways to “survey” the chapter so as to become familiar with the organization and main ideas they were about to encounter while reading. During the second pass, students were taught to “size-up” the chapter by skimming the text looking for factual details they would encounter, novel vocabulary and various details from textual cues. Finally, the students were taught “sort-out” the information that was presented in the text, answering questions and searching back through the text for answers. Participants were taught each strategy and each pass through a meticulous, ten step, teacher-led training procedure to ensure each student learned each strategy to mastery. Dramatic and sustained results were observed in their multiple baseline research, where each student immediately showed improvement on their ability to answer questions related to the texts and on their ability to verbalize what strategies

were expected of them. Importantly for struggling learners, all students showed dramatic improvement on ability level tasks and grade level tasks.

Many studies on multiple strategy instruction incorporate an instructional tool that utilizes peer-interactions to increase student dialogue and scaffold “teacher talk” to “kid-talk (Klingner, Vaughn, & Schumm, 1998).” Most commonly known as reciprocal teaching (Palinscar & Brown, 1984), this approach entails teaching students a set of cognitive strategies (typically making predictions, generating questions, using clarification strategies, and summarizing) through teacher modeling of each strategy and subsequently through student modeling for one another. The gradual release of each strategy from teacher to students is done strategically as the teacher actively supports the practice dialogue by rephrasing and expanding upon the students responses, essentially shaping the ideas until eventually the responsibility of active comprehension is in the control of the students. Theoretically, reciprocal teaching “provides active rehearsal in inner dialogue and self talk that the reader covertly uses during comprehension” and “promotes the internalization of these strategies and the development of self-regulation (Englert & Mariage, 1990).”

Similar to reciprocal teaching, collaborative strategic reading (CSR; Klingner, Vaughn & Schumm, 1998; Klingner & Vaughn, 1998) are based on the same principles to provide students with scaffolded practice in comprehension strategies, but CSR relies more on small groups working together and less teacher-led discussions. The strategies taught in CSR include Previewing the text, the Click and Clunk clarification technique, Get the Gist strategies for extracting the main idea and a Wrap Up. Students are directly taught each element and then sent to work through texts in small groups. Instead of

facilitating the discussion, the teacher's role shifts to supervising the small groups and assisting only when necessary.

Providing students opportunities to practice skills what they are taught is an important feature of any instructional plan. Reciprocal teaching and CSR provide just that, organized scaffolding for practice of a complex set of cognitive strategies. There is strong evidence to suggest that providing students with multiple strategy instruction has a positive effect on their performance on experimenter designed measures and has more promise in producing learning that will generalize to other measures of reading achievement (Gersten, et al., 2001). Importantly, Rosenshine and Meister (1994) reviewed the extensive research on reciprocal teaching strategies and reported evidence of improvement on both experimenter designed tests for students who were provided such strategy instruction when compared to those who were not, as well as significant improvement on standardized test of reading. These results show promise for the generalization of skills taught using student's to scaffold their own opportunities for practice.

Graphic Organizers and Text Structures. The members of the NRP identified one final instructional tool for teaching students to apply cognitive strategies to devise meaning from text. Graphic organizers are commonly used for students to organize their ideas as pictorially so as to visually depict relationships of elements in a text. Very often graphic organizers are used as a scaffolding tool for students to record their responses to prediction, clarification, questioning and summarization tasks. These instructional tools have been found to help in aiding students to work independently on strategies that have been taught to them, focus on what they read and assist in the organizing written

summaries. Mostly, graphic organizers aid students in remembering content of what they read.

Finally, there has been some evidence that teaching students text structures and passage organization will assist in comprehension. Students who are aware of text structure as they read approach the content with a plan of action and are better at organizing the information in the text in a meaningful way (Meyer, Brandt & Bluth, 1980). Wong and Wilson (1984) found significant results when they taught passage organization to children with learning disabilities. The researchers found the learning disabled students in their sample were unable to identify disorganized sentences and passages, and less able to reorganize the words in the sentences into meaningful units. Importantly, the researchers found training struggling students to recognize relationships between sentences in a paragraph was not only simple, but also improved their ability to sort disorganized sentences into meaningful passages. While instruction in common sentence structure, genres and text structures appears to benefit struggling readers, the same training appears to be superfluous for typical students (NRP, 2000).

Summarizing the Evidence in Reading Comprehension Instruction

Thus far, research has provided educators with some insight into probable strategies students employ to aid in comprehension and ways to address these strategies through direct instruction. Of all the ideas that have been tried and tested, evidence seems to identify three specific strategies, cognitive awareness, question generation, and summarization. Strikingly, these three strategies are rooted in solid theories from a cognitive psychological perspective. Strategy instruction in comprehension awareness is rooted in the theory of the struggling student as an inactive learner who only needs to be

taught to recognize when comprehension breaks down and how to stop and fix the misunderstanding. Summarization strategy instruction stems directly from Kintsch and van Dijk's rule-based cognitive theory which purports a reader's prior knowledge of both structural elements (microstructures) and content (macrostructures) is the main vehicle by which meaning is constructed. Question-generation instruction integrates both theories. Teaching students to generate questions while they read encourages active attention to the content of the text, however, original research on question generation (Singer, 1978) was designed according the premises of Kintsch's theory that meaning is constructed as an interaction between resources the reader brings to the activity and characteristics of the text. The original goal was to train students in elements of the story structure so they would generate questions with more student-based information than text-dependent information.

It is important to recognize the research on comprehension instruction is grounded in solid theory, and there exists a growing body of evidence to support direct instruction in these three comprehension strategies, however, it is equally important to examine the quality of the research and acknowledge the existing flaws so as to accurately interpret the evidence and its implications for practice before the practices are identified as "scientifically-based." Fortunately, federal law, through the Reading Excellence Act (1998) has provided guidelines for doing just that. Scientifically-based research has applied rigorous, systematic and objective procedures that employ systematic empirical methods, involves rigorous data analyses that are adequate to justify the general conclusions drawn, relies on reliable and valid measurements and observational methods, and have been accepted by a peer-reviewed journal or approved by a panel of

independent experts through a comparably rigorous, objective, and scientific review.
(“Reading and Literacy Grants to State Educational Agencies,” Title II (C) Sec. 2252 (5)
[20 U.S.C. §6661a])

Despite the claims of the NRP, that they present evidence of instructional practices that “appear to have a firm scientific basis for concluding that they improve comprehension in normal readers (NRP, 2000, p. 4-42), later in the body of the text the panel acknowledge the various flaws in the reading comprehension research that would indicate unresolved results. Importantly, reading comprehension research lacks reliable and valid measurements and observational methods by which to judge experimental results.

In a 1989 methodological analysis designed to examine the empirical methods used in reading comprehension research, Lysynchuk and her colleagues found most reading comprehension research falls short of the standards that define scientifically-based practices. They observed the majority of the studies did not control the research design with rigorous, objective procedures so as to draw valid conclusions and rule out other possible explanations for the studies’ results. Random assignment to students between the control and experimental conditions were observed only 64% of the time. In almost 30% of the studies, the control group was exposed to different instructional materials as well as different instruction, making it impossible to conclude whether it was materials or instruction that led to group differences. Similarly, when time on task was reported, it was not clear whether students spent the same amount of time, thus conclusions about the efficiency of instructional practices can not be drawn.

Lysynchuk et al.'s (1989) analysis also revealed most often the data analyses that were used could not adequately justify the general conclusions drawn from the research. In more than 80% of the studies, inappropriate units of measure were analyzed, where the unit of analysis was not the same as the unit of treatment. Although about 90% of the studies did use otherwise appropriate statistical procedures based on probability theory to control Type I error rates.

Finally, most relevant to this research, Lysynchuk et al. (1989) examined the quality of the dependent variables used in the reading comprehension instruction research. They found that only 35% of the dependent measures were known to be reliable, and they did not evaluate the measures for their validity as outcome measures indicative of reading comprehension. An examination of the studies evaluated in the NRP evaluation reveals experimenters designed tests that directly assessed comprehension of each specific reading task was the most common dependent variable (Appendix A). Often these tasks involved question-answer tests and recalls and summaries. Direct observations of the students implementing the strategies that were taught to them was also a common dependent variable. Very few studies evaluated strategy instruction with a far transfer measure, when it was done, standardized test of reading comprehension and content area achievement was evaluated. In examining the studies of external validity, Lysynchuk and her colleagues found only 24% of the studies evaluated delayed effects. Importantly, to evaluate the delayed effects and generalizability of reading comprehension strategy instruction, it is important to have a consensus on what that general outcome is.

The NRP (2000) concluded "the empirical evidence reviewed favors the conclusion that teaching of a variety of reading comprehension strategies leads to the increased learning of the strategies, to specific transfer of learning, to increased memory and understanding of new passages, and, in some cases, to general improvements in comprehension (p 4-52)." Despite the massive efforts of the NRP to delineate effective practices for teaching reading comprehension, the RAND study group stated, there currently exists little evidence regarding what type of instruction will promote reading comprehension outcomes. The RRSg purports more work must be done to investigate how best instructional time should be spent if the goal is improved comprehension. After reviewing all the evidence, they still assert there is no conclusive evidence that instruction in fluent word recognition strategies, vocabulary, or explicit teaching of reading strategies will produce meaning gains in reading comprehension for students. Primarily, these questions can not be answered until meaningful outcome measures of reading comprehension are developed and then evaluated for their reliability, validity of near transfer and ideally validity of far transfer. Now is the time to focus on developing assessments based on good theory to judge the effectiveness of reading comprehension instruction on a group level and formative measures to evaluate instruction on an individual level.

Assessing Reading Comprehension

A difficulty in discussing consequences [of reading comprehension instruction] in any great detail, is that only very limited of assessments of reading comprehension are available....[U]ntil comprehension measures expand to reflect an underlying theory that acknowledges a variety of possible consequences, both immediate and long-term, we will be severely hampered in our capacities to engage in excellent research on this topic. (RAND, 2002, p 110)

Reading comprehension involves the orchestration of word recognition strategies, vocabulary awareness, prior knowledge and a variety of cognitive strategies. Assessment of the cognitive activities involved in reading comprehension must be broad enough to reflect its complexity yet specific enough to accurately reflect meaningful changes as a result of learning and skill development. Assessment tools must be sensitive to the identification of differential abilities between groups of students as a result of instruction, specifically for the empirical evaluation of effective instructional practices. The availability of valid and reliable measures is also important to assess changes within individual students as a result of instructional changes.

The members of the RRSg (2002) assert a research agenda for reading comprehension instruction and subsequent improved reading comprehension outcomes for students depends upon a robust assessment system that reflects the dynamic nature of reading comprehension across texts, purpose and individuals. Such a system should be robust so as to reflect the development of reading comprehension abilities as observed through engagement with text, knowledge acquisition and the application of new information. Unfortunately, most available assessment tools rely too heavily on the memory of text content, vocabulary and word reading ability and are designed with too little attention to theories of reading comprehension.

Assessing the effectiveness of reading comprehension instructional strategies on a group level and for individual students requires a consensus regarding what specific outcomes, if observed, would reflect sufficient comprehension. Using current theoretical frameworks to identify observable indicators of successful comprehension provides insight into two possible outcomes, the process and the product. It is possible to design

assessments around the process of comprehension, where students effectively react to text in specific ways so as to actively construct meaning. Secondly, it may be beneficial to construct measures of comprehension around the product of successful reading, where the student constructs a personal representation of the text, which is essentially meaning.

The RRSg (2002) suggests research for the development of reading comprehension assessment tool should focus on the capacity to reflect authentic product of reading comprehension as a result of good instruction. They stress the importance of a predictable consistency between assessments and the processes involved in comprehension and thus tools that are developmentally sensitive to these processes. Assessments should have both the capacity to distinguish between poor and good comprehenders so as to target needy students and the ability to identify subtypes of poor comprehenders so as to design strategic instruction. Importantly, assessments must be sensitive to instruction so as to reflect learning when it occurs and modify instruction when necessary. Finally all assessment measure should be grounded in a solid foundation in psychometrics (RAND, 2002, pp. 56-57.)

The primary purpose of good assessments should be to provide teachers with information that will have a positive impact on their academic performance. Curriculum-based assessment (CBA) and curriculum-based measurement (CBM) procedures have played an important role in improving academic outcomes for children. Though related and equally informative, CBA and CBM have two important distinctions (Fuchs & Deno, 1991). CBA has more commonly been used in classrooms and is commonly referred to as specific subskill mastery. Using CBA, teachers are able to assess their students' degree of mastery according to each lesson in the scope and sequence of the

curriculum. CBA provides teachers with important information about what skills their students may have not yet mastered. CBM is better described as measuring general outcomes, where general outcomes are the eventual level of proficiency one would expect from students in a general skill area. Whereas, CBA requires a process of evaluation including assessments and observations, CBM relies upon standardized methods for measuring students' levels of performance in specific academic areas across the school year. CBM as it was originally conceived may provide a valuable framework for developing quantitative assessment tools for reading comprehension, while CBA may provide the qualitative information necessary for instructional planning.

Originally designed in the late 1970's as part of Special Education instructional model where repeated measurement data was used to evaluate the effectiveness of instruction and provide evidence to adjust instruction accordingly (Deno, 2003), CBM has been identified as valuable in improving outcomes for students receiving Special Education services as well as students participating in the general curriculum (Fuchs, 1986). CBM assessment procedures produce a database of students performance data in specific content areas, which graphed over time provides information about students development in the skill area as a result of instruction, thus allowing teachers to change and modify their instruction based on standard decision rules.

Measures used for CBM must be efficient and easy to use so as to collect a database to adequately reflect student learning over time. In addition, they must provide observations of authentic student learning sensitive to change as a result of instruction. Finally, they must be standard tasks that are reliable and validly reflect the construct being measured. Currently, Maze (MZ) is commonly used as the standard reading

comprehension CBM task, but is it worthwhile to examine whether the process or product of reading comprehension, as complex as it is, can be observed using CBM procedures based on the rigorous criteria.

In the massive body of reading comprehension strategy instruction research, most hypotheses could be summarized as one of two things. "If students constructed a particular mental representation of texts, or reacted to the text in a particular way, then comprehension and long term memory of the text would improve (Pressley, Brown, El-Dinary & Afflerbach, 1995)." From a product perspective the observation of reading comprehension would be observed as a meaning unit that does not represent purely the orthographic structures of a text, but a newly constructed meaning representation (Royer & Cunningham, 1981; Kinstch & van Dijk, 1978). From a process perspective, the observation of effective reading comprehension would reflect students' active awareness and control over their own understanding and memory for the text (Markman, 1979; & Torgesen, 1977, 1982).

Royer and Cunningham (1981) best summarize the production-based theories of reading comprehension with the "minimal principle of reading comprehension," which simply states, "the act of comprehension entails an interaction between an incoming linguistic message and the comprehender's world knowledge (p. 188)." For example, Kinstch and van Dijk (1978) explain that meaning begins at the word level where meaning is constructed as it related to word meaning. Subsequently, meaning is then derived at the phrase level where the words in the phrase are recognized lexically, then the meanings of each word constructs phrase-level meaning, and so on and so forth. Additionally, Schank (1972) proposed the theory that once an individual accesses the

lexical input of words in a text, meaning is subsequently constructed by sorting the words according to their functions in the text. Nominal words are typically nouns and stand alone as the object in the text that begins the picture, action words, often the verb, describe what the nominal concept is doing. Modifiers then describe the nominal concept and the action concept. Meaning is thus constructed by a series of processors that perceive the syntactic input, categorize the concepts, relates the concepts and then stores the information as a meaning unit.

The most common proxies used as dependent variables in the reading comprehension research to measure the product of new meaning constructed as a result of effective comprehension processes are summaries and question-answer tests. Neither of these two possible measures of meaning construction has been done according to standard testing procedures, and both are less amenable to being efficiently useful CBM measures. However, the sentence verification technique (SVT) is specifically designed to observe the variability in comprehension of students who remember the lexical cues of the text versus the meaning in the text (Royer, 2001). Importantly, the test is efficient to administer according to standard testing procedures and can be used with any curriculum (Royer, 2001).

Because these theories describe interpreting the lexical input as the first step in meaning construction, that is then reciprocally honed through interactions between understanding and higher levels of orthographic information, it makes sense the measurement of comprehension under these theoretical frameworks includes an observation of the newly constructed meaning as well as the fluidity with which the words were read, easily measured with an oral reading fluency task (ORF).

Theories of comprehension monitoring or metacognitive awareness, describe the general outcome of comprehension as mature thought. Evidence demonstrating developmental differences in comprehension monitoring strategies, supports the hypothesis that awareness of one's own thinking processes may indeed be developmental, and teaching students' strategies for self-monitoring illustrates metacognitive awareness can also be taught (Wong, 1980).

Because metacognitive awareness is difficult to define as general outcome criterion, observations of appropriate interactions with text to construct meaning may best be evaluated using an assessment process, such as CBA. However, it is possible the Maze (MZ) CBM task assesses students ability to make clarifications during reading, a definite proxy for self-correction strategies. More commonly, metacognition is determined by evaluating whether students apply strategies that were taught to them. Howell and Nolet (2000) describe the assessment of reading comprehension as an evaluative process, which hierarchically rules out enabling skills, such as decoding, vocabulary, syntactic awareness and prior knowledge, and then explores how the student employs comprehension strategies, including monitoring meaning, selective attention to the reading task, connecting text to prior knowledge and clarification of confusions.

Informal reading inventories have also been used that provide researchers qualitative information about students' instructional levels in reading accuracy, fluency and comprehension, and the comprehension questions allow for an evaluation of the various types of comprehension failures students make (Dewitz & Dewitz, 2003). Finally, think aloud observations and verbal protocols have been commonly used in reading comprehension research, measuring innumerable possibilities in cognitive

processing. In fact, Pressley and Afflerbach (1995) compiled over 40 verbal protocols, sorting the typical observation criteria of all possible cognitive processes that have been measured into specific processing criteria. This work may provide a good start for a standard observation battery to measure the development of cognitive awareness over time.

In its 2001 research agenda the RRSg called for educational researchers to focus their attention on the development of valid and reliable reading comprehension assessment tools with the capacity to reflect authentic outcomes and in both the outcomes and the processes involved in comprehension. They suggested assessments be investigated that are sensitive to development, instruction and variability between students. Finally they called for measures that can accurately identify variability within students and identify subtype of poor comprehenders, thus providing teacher with valid instructional information. These criteria suggest comprehension assessments that reflect the standards of CBM and are designed for efficiency to allow for the collection of a database of information about the progress of reading comprehension skills, would benefit analysis of both research on effective practice and for the assessment of individual students.

Implications for Practice

While further work must be done to investigate reading comprehension research, students must still learn how to read with purpose and effectively construct meaning. Thus teachers must continue to teach students comprehension strategies that provide them skills to extract meaning from any text they might encounter. In the absence of direct evidence to support one instructional strategy over another, to help prioritize

instructional time and to differentiate instruction, educators can still be empowered to provide their students effective instructional practices by relying on reason-based practice that converge with the science literature. (Stanovich & Stanovich, 2003). For the first time in thirty years, the National Assessment of Educational Progress (NAEP, 2004) reported nine year old students scored higher in reading than in any other assessment year. It is time to use evidence and reason to continue the upward trend in literacy outcomes for American students.

It is time to provide all students with a strong foundation in phonemic awareness, with explicit instruction in systematic phonics to develop automatic word recognition strategies that result in fluent reading. It is time to ensure every student has access to direct instruction in specific strategies to aid in reading comprehension, such as clarification, summarization and question generation techniques. These strategies are grounded in solid theoretical foundations with a growing body of evidence to suggest they would be the most advantageous approaches for teaching students to comprehend text. Although research based on rigorous scientific methods, using valid and reliable measures of reading comprehension are necessary to provide direct evidence supporting reading comprehension is necessary before widespread proliferation of specific instructional advice, reason as well as research should guide practice the a maturing professional field of education.

CHAPTER 3

METHOD

Participants

An a priori power analysis was conducted to determine the number of participants necessary to satisfy the assumptions of a multiple regression analysis. With an α of .05 and a medium effect size of .15, results of the analysis indicated that a sample size of 125 would result in power of approximately .80.

Prior to seeking parental permission, approval was granted by the Superintendent of the school district and the principals of participating elementary schools. Because the tests involved were an extension of the typical assessments conducted within the district, parents could refuse their child's participation by returning the consent form (see Appendix B). Before beginning testing, each student was informed of the purpose of the study and what their participation would entail. At that time, students were able to opt out of testing if they did not want to participate. Two students chose not to participate, and four parents requested their children not participate.

Settings

Data were collected from 111 fourth grade students from two Western Massachusetts elementary schools. Both schools participate in the Massachusetts Reading First Initiative. The purpose of this initiative is to support the implementation of high-quality, scientifically-proven methods of reading instruction in K-3 classrooms with the goal of preventing reading difficulties so all students will be proficient readers by third grade. Thus both schools utilize early literacy instruction as delineated by the NRP, with direct and systematic instruction focusing on phonemic awareness, phonics, fluency,

vocabulary and reading comprehension. Participating schools have designed their reading instruction around a three-tiered model, in which the core curriculum serves the majority of the students. Strategic and science-based interventions are integrated into the instruction for students who are identified as needing more specific skill instruction, and very intensive interventions are utilized for students who are identified as the most needy. Formative assessments are used to determine students' needs within this continuum of instruction as well as their rates of progress in relation to the instruction they receive.

Materials

Independent Variables

Individually Administered Measures

Oral Reading Fluency (ORF). Students were presented with three fourth grade reading passages consisting of approximately 350 words drawn from Edformation's Standard Reading Assessment Passages (Shinn & Shinn, 2002). In accordance to the standardized instructions, the students were asked to do their best reading. For one-minute, each student read aloud while the examiner read along on a separate scoring passage, marking mispronunciations, word substitutions, omissions, and 3-second pauses as errors. At the end of one-minute, a bracket was inserted behind the last word read. This procedure was repeated for three passage read alouds. The ORF score was calculated by counting the words read correctly on each passage and then taking the median of the three scores. ORF has been demonstrated as a reliable assessment tool with observed test-retest reliability between .82 and .97 (Marston, 1989, & Shinn & Shinn, 2002) and an inter-rater reliability of .99 (Shinn & Shinn, 2002).

Retell Fluency (RTF). RTF was collected concurrently with ORF according to the standardized DIBELS directions (Kaminski & Good, 2003). After reading each passage aloud, the students were asked to retell all they could remember about what they had just read. Students who hesitated for three seconds or more were given a prompt to tell everything they could remember. Using a numeric tracking grid, the examiner counted the words the student retold, excluding nonwords and repetitions. At the end of the one-minute period or after a five-second hesitation or irrelevant track, the last number in the grid was circled. This circled number is the total words recalled. The score for RTF was derived from counting total words recalled for each passage and taking the median score from the three retells. Inter-rater reliability for RTF has been observed between .96 and .98 while alternate form reliability has ranged between .47 and .70 and test-retest between .58 and .78 (Manchester et al., 2004).

Group Administered Measures

Sentence Verification Technique (SVT). SVT is designed to evaluate how students extract meaning from text. SVT consists of four reading passages. Each passage contains 12 sentences that are to be read silently, followed by a series of 16 sentences, some having the same meaning as sentences in the passage while others have different meanings. Students are asked to judge which of the 16 sentences have the same meaning as the sentences in the passage.

Each student was provided one SVT testing packet including four passages each followed by their 16 respective test sentences. The first passage and corresponding test sentences were of third grade difficulty. The second and third were of fourth grade difficulty, and the fourth passage was based on a fifth grade readability.

The students were guided through the instructions along with a practice example. First they were told the purpose of the test is to see how well they understand what they read, followed by general directions for taking the test. Then they were asked to read the practice passage carefully so they could answer questions when they are done. The students were instructed to answer the test questions by marking "yes" next to questions that meant the same thing as a sentence in the story they just read and "no" next to those sentences that meant something different. Finally, the students were guided through the practice test questions with a demonstration of how the correct answers were derived.

After the guided example, students were read the remaining standard directions, instructing them to read each story carefully and turn the page to answer the questions without turning back to the story. When they finished one story, they moved on to the next until they had read all four passages and answered all corresponding questions. Each student was given 30 minutes to complete this test. The score for SVT was derived by counting sentences marked correctly.

Reliability for SVT has varied between .5 to .9, depending upon the number of passages and questions provided to students. SVT assessments that employ four passages with 64 test sentences have produced reliability estimates of .70 to .80 (Royer, 2001).

SVT appears to be a valid measure of reading comprehension. Research suggests SVT correlates with teacher judgement regarding student comprehension skills (Royer & Carlo, 1991). Royer (2001) also reports a positive relationship between SVT and the reading comprehension subtest of the Iowa Test of Basic Skills (ITBS; $r=.73$). SVT also correlates with other standardized tests that require students to read and extract meaning

from text; specifically SVT correlates with the science and social studies subtests of the ITBS at .72 and .65, respectively.

Written Retell (WRT). WRT is a procedure that measures students' ability to attend to and recall details and main ideas of what they read. It requires students to read a passage for five minutes and then, with the passage withdrawn, write the recall for a five-minute period. In this study, students will be presented with a 750-word passage taken from the Houghton-Mifflin Grade Four curriculum, and asked to read the passage silently for five minutes. At the end of the five minutes, students were asked to write all they could remember about what they just read and were given five minutes to write their retell, with a prompt to keep writing at each minute mark. Two scores were derived from written retell: (a) total words written and (b) unique content words written. Content words were operationally defined as distinct proper and common nouns, verbs, adjectives, and adverbs contained in the passage or synonymous with those in the passage. A variation of this WRT procedure has demonstrated criterion validity compared with standard measures of reading comprehension; the correlation coefficients have ranged from .76 and .81 (Fuchs, Fuchs, & Maxwell, 1988).

Maze (MZ). The MZ task theoretically measures students' vocabulary knowledge, syntactic understanding and ability to employ active reading strategies for meaning monitoring while they read (Howell & Nolet, 2000). In this study, students were given three fourth grade passages in which every seventh word was replaced by a multiple-choice selection of three word choices. The choices included the correct word, a syntactically correct word that does not make sense in the sentence and a third distractor that neither syntactically nor semantically makes sense in the context of the sentence.

Students were asked to read each passage silently, and when they came to a group of three bold words, choose the word that made the most sense. Students were given three minutes to work on each passage. When the three minutes elapsed, the administrator stopped the students from working and guided them to the next passage to begin working on. Scores were derived from simply counting the number of correctly circled words for each passage and then taking the median score. Alternate form reliability for MZ has been observed between .62 and .93 (Guthrie, 1973). Test-retest reliability has been reported at .83 (Manchester et al., 2004).

Criterion Measures

The Group Reading Assessment and Diagnostic Evaluation (GRADE). The GRADE (Williams, 2004) is a standardized test designed to assess reading skill development from pre-reading early literacy skills through decoding and comprehension skills. It is organized by grade levels and divided into subtests reflecting the literacy skills at each level. The subtests that make up the fourth grade level of the GRADE include Sentence Comprehension, Passage Comprehension, Listening Comprehension and Vocabulary. This test elicits both a standardized comprehension composite that includes the sentence and passage comprehension subtests as well as a standard score for the total test, including the two reading comprehension subtests and the vocabulary subtest. The criterion measure that was used for this research was the overall reading proficiency score elicited by these three subtests.

The purpose of the Sentence Comprehension subtest is to determine whether students can construct the meaning of sentences as one whole idea unit. In the fourth grade level of this subtest, students are presented with 19 sentences in isolation. Each

sentence is missing one word. The students must determine the single word that is missing from a list of four to five choices listed below the sentence. The distractors were carefully selected to be plausible options that do not make sense in the context of the sentence and are not so ambiguous as to trick the examinees. Students who do well on the Sentence Comprehension subtest demonstrate their proficiency in reading and comprehending isolated sentences as a whole thought regardless of the grammatics or semantics of the sentence structure. This subtest has reported a split-half reliability of .92 (Williams, 2001).

In the Passage Comprehension subtest, students are presented with one short, two medium and three long passages. Each passage is followed by between three and five multiple-choice questions for a total of 28 questions in Level 4 of the GRADE. The purpose of this subtest is to measure students' ability to use metacognitive comprehension strategies while reading, including questioning, clarifying, summarizing and predicting. Thus questions in this section are specifically related to these four comprehension strategies. The passages include both excerpts from authentic literature and synthetic passages designed to target specific reading skills so that analysis of test performance has instructional utility. The split-half reliability of .94 for this subtest has been observed (Williams, 2001). Students who do well on the Reading Comprehension subtest demonstrate the ability to read and comprehend different text, topics and genres through the use of metacognitive comprehension strategies.

Vocabulary subtest in Level 4 of the GRADE assesses students' knowledge of words appropriate to their own grade level. A poor performance on this subtest suggests the student might not have the adequate vocabulary to comprehend text at the specific

grade level. Level 4 consists of 35 test items in which the students are presented with a phrase or short sentence in one boldfaced word. The students must determine the correct meaning of the boldfaced word from a list of four possible answers listed below the item. The split-half reliability for the vocabulary subtests was .94 (Williams, 2001).

Vocabulary knowledge is both an enabling skill for comprehension and reciprocally is strengthened through proficient reading comprehension (Howell & Nolet, 2000).

As part of the general education practices in the participating schools, each student was administered the Level four of the GRADE at the beginning of the year and was assessed once again in the Spring. Reliability for each test level was .90 or greater. Test-re-test reliability was between .77 and .98 and alternate form reliability has been reported between .81 and .94. Overall, concurrent validity between the GRADE and the Iowa Test of Basic Skills for reading ability has ranged between .69 and .83 (Williams, 2001).

Massachusetts Comprehensive Assessment System (MCAS): The MCAS is an untimed criterion-referenced test designed specifically for the state of Massachusetts. The test questions are rigorously based on the Massachusetts curriculum frameworks, reflecting the high academic standards expected of the students and educators in the state. Committees of teachers throughout the state developed questions in direct relation to the grade level standards.

The English Language Arts section of the MCAS was used as the outcome measure for overall academic success. This section of the state assessment consists of two components, language and literature, and composition. The language and literature section is designed to assess students' skills with the structure and dynamics of the

English language and their ability to respond to spoken and written language. The composition section assesses students' writing in terms of clarity, focus, organization and language usage as well as grammatics and punctuation.

The English Language Arts assessment for fourth grade consists of 36 multiple choice and short answer questions, eight open-ended questions and two writing prompts. Non-objective test items are scored on a standardized rubric specific for each question. Overall scores are categorized into four levels: Advanced, Proficient, Needs Improvement and Warning.

Procedures

Training Data Collectors

A team of 10 school psychology graduate students were trained in the standard assessment procedures for ORF and RTF. They were trained using concurrent testing procedures in which trainees were first directly instructed in the standard administration and scoring procedures. Immediately following instruction, the trainees practice scoring along with a pre-recorded video of a student's orally reading and retelling stories similar to what they could expect to encounter during the natural testing conditions. Examiners practiced concurrently until they reached an acceptable consistency criterion of within two points of each other's final score (Good & Kaminski, 2003) for three consecutive tests.

A second examiner was trained in the scoring procedures of the WRT. Once again the trainee was provided direct instruction in the scoring procedures. The trainee was provided a comprehensive list on all unique content words contained in the original story with each word divided into columns according to its part of speech. Following the

training, the second examiner practiced scoring five WRT tests where the two examiners quickly reached consistency scoring exactly the same on two tests and within two points of one another on three of the tests.

Interrater Agreement and Procedural Integrity

Fidelity in the administration and scoring of the individual assessments was checked on approximately 20% of the sample using an Assessment Integrity Checklist and a synchronous testing check. To assess the consistency between examiners during individualized testing of ORF and RTF, one out of every five administrations was scored concurrently, with two examiners scoring along as one student read and retold, and evaluated according to the DIBELS Assessment Integrity Check (Appendix C; Good & Kaminski, 2002). Initially, the agreement between the examiner and integrity assessor was calculated using the percentage agreement formula: $(\text{Agreements} / (\text{Agreements} + \text{Disagreements})) \times 100$. However, results of the testing revealed the validity of the two-point criterion for agreement published by the test developers appears questionable and the criterion itself, and unreasonable expectation for a band of error. Thus, correlations between the observer scores were computed, as well as scatterplots, were used to evaluate the reliability of the data in order to determine if they were indeed reliable to include in the analyses.

Data Collection

Individualized administration. Oral Reading Fluency and Retell Fluency was collected in concert with one another according to the standardized DIBELS instructions. Data were collected individually, with each student reading a standard fourth grade passage aloud for one-minute and then retelling what they read for another minute. Each

student read and retold three passages with a total of 6 minutes of individualized testing for each student. The median of words read and retold correctly were recorded for the ORF and RTF scores, respectively.

Group Administration. SVT, WRT and Maze were administered in a group format to each of the six classes of students. Not including the time it took to read the standard directions and practice questions, students were given a total of 30 minutes to complete the SVT test. Total testing time for WRT was 10 minutes, where five minutes were allowed for the silent passage read, immediately followed by five minutes to write the retell. Maze, lasted a total of nine minutes, three minutes per passage. Total group testing time was approximately 50 minutes per class, excluding the time it took to read all of the directions and administer the practice items. Administration of the tests were counterbalanced by class, controlling for placement and ordering effects of the testing procedures (Table 3.1).

Table 3.1:
The counterbalancing design of group administered tests.

	Test 1	Test 2	Test3
Class 1	MZ	WRT	SVT
Class 2	WRT	MZ	SVT
Class 3	SVT	WRT	MZ
Class 4	MZ	SVT	WRT
Class 5	SVT	MZ	WRT
Class 6	WRT	SVT	MZ

Criterion Tests. GRADE data were collected in the Fall and Spring for each student as part of the general schoolwide reading program. GRADE data were collected by school staff, and the researcher received the results for each fourth grade student from the Spring GRADE Comprehension Composite standard score as soon as they were available. MCAS is conducted statewide in late March and early April, although the

scores are not released until the Fall. When the results were reported, the Reading Specialists from each school compiled the scores from only fourth graders who participated in the study and sent them to the researcher for analysis.

Teacher Rating. In addition to the measures of comprehension, teachers were asked to rate their perceptions of their students reading comprehension ability. They were given a grid with their students' names along the left side and the numbers 1 through 9 alongside each name. The goal was to have them rate their students in the form of a normal distribution; the numbers 1 through 9 represented the stanines of their class. As a school that has participated in the Reading First Initiative, the teachers were aware of stanines as a way to interpret score. However individually, the teachers were directly instructed in how to rate their children on the stanine scale.

Data Analysis

Screening

Prior to the regression analysis, data were evaluated for outliers, normality, linearity and homoscedasticity as they pertain to the assumptions underlying linear regression. Screening for outliers was done by evaluating Mahalanobis distances for each multiple regression analysis. While it is not feasible to assess the normality of the multivariate distribution, Q-Q plots were evaluated for the normality of the univariate distributions. The statistics for each indicator of normality was also examined for any non-normal indicators.

Examining Relationships between Variables

A correlation matrix was computed to compare the relationship between the dependent variable, GRADE, and each of the independent variables, the formative

assessment measures, as well as the relationships between each of the independent variables. This correlational analysis revealed an informative picture of the function of each assessment within the regression model as well as the contribution of each formative assessment measure to the construct of reading comprehension. The correlational analysis was also used to evaluate whether multicollinearity existed for any of the independent variables in the model.

Exploring Variability in Reading Comprehension

Initially four multiple regression equations were computed to determine whether any of the independent variables measuring reading comprehension significantly predicted reading ability above and beyond what is explained by ORF. This significance of each predictor was evaluated, as well the partial correlation to examine the degree of the relationship between reading comprehension and each formative assessment measure. *R*-squared was calculated to understand the proportion of the variance accounted for by each predictor variable.

Deriving a Model of Reading Ability

A multiple regression was computed to evaluate the significance of a comprehensive model of reading competence, where ORF, MZ, SVT, RTF, and WRT were used to predict reading competency as observed by performance on the GRADE. The specified model is represented by:

$$GRADE_i = \beta_0 + \beta_1 (ORF_i) + \beta_2 (MAZE_i) + \beta_3 (WRT_i) + \beta_4 (RTF_i) + \beta_5 (SVT_i) + \varepsilon_i.$$

To determine which independent variable significantly predicted reading ability above and beyond ORF, the Holm Sequentially Rejective Procedure Procedure (1979) was applied to the results of the multiple regression analysis. This procedure is designed

to evaluate the significance of each predictor through a series of stepwise comparisons. The Holm procedure was chosen because of its ability to control the familywise error Type I error rate, yet maintain a meaningful degree of power so as to determine which measures significantly predict reading ability despite controlling for all the variability accounted for by the other predictors. Additionally, this analysis was used to reveal the most parsimonious model to best predict performance on the GRADE.

Predicting Academic Achievement

Once the previous analysis that explored the relationship between the independent variables and the criterion, reading comprehension (GRADE), had been explored, a logistic regression equation was constructed that included the significant partial regression coefficients so as to evaluate how well the significant predictors for GRADE could classify students as passing or failing the fourth grade English Language Arts section of the Massachusetts Comprehensive Assessment System. Using this analysis, it was possible to determine whether a relationship existed between performance on the state assessment classification and the set of reading measures in the model.

Finally, in the vein of exploratory research, a multiple linear regression was performed with the same predictors as the previous logistic regression and the MCAS ELA score as the dependent variable. This analysis was performed to examine how much information was lost by dichotomizing the ELA scores. In addition, this analysis helps determine whether a model of reading, using both a measure of reading fluency and a measure of comprehension, would predict academic success based on state standards of grade level achievement. For assessment measures to be useful, they must be reliable, have construct validity and predictive utility.

CHAPTER 4

RESULTS OF THE ANALYSES

Summary of Purpose and Method

The purpose of this study was to examine the incremental and predictive validity of four formative measures of reading comprehension: MZ, RTF, WRT and SVT.

First, it was hypothesized that, due to the many subskills subsumed within the process of reading, formative assessment measures of reading comprehension will contribute more to our understanding of reading ability than simply the measure of oral reading fluency.

Four multiple regressions were computed to test this hypothesis, where each analysis included the measure GRADE as the dependent variable and ORF as an independent variable.

Furthermore, it was hypothesized that valid formative measures of reading comprehension can be modeled in a meaningful way so as to predict performance on measures of overall academic success. A sequential multiple regression was computed to determine which of the measures of reading comprehension appeared most powerful in predicting overall reading ability in addition to ORF. In this analysis, ORF was built into the model first and then reading comprehension measures were added to the model as a block. Finally, with the results of the sequential multiple regression, the single most powerful predictor of reading comprehension was included in a logistic regression model with ORF to predict academic success, represented by the MCAS English Language Arts examination as the dependent variable.

Screening for Underlying Assumptions

Data were screened to determine adequacy of the analysis to the data set before conducting the initial regression analyses. A total of 111 students were included in the sample. Three students were not included because of missing data. Two students chose to stop participating during the testing session. Four students did not participate because their parents chose not to consent to the testing. One student chose to participate, but his reading ability was so poor, his Special Education teacher had to read him the tests; thus his scores were not included in the analysis. Finally, one outlier was removed from the logistic regression analysis because his scores did not represent the sample population, where his formative measures were extremely low yet he passed the MCAS. In further examination of the case, it was discovered he had a test accommodation where the MCAS was read for him.

While it is not possible to assess the normality of the multivariate distribution used for these analyses, Q-Q plots were computed for the univariate distributions of each measure to examine univariate normality (see Appendix D). Most of the measures were approximately normally distributed, except perhaps the distribution for SVT where the kurtosis (1.34) was inflated, indicating non-normality with the extreme cases. Further examination of the distribution revealed the cases were negatively skewed. While this is an indication of a non-normal distribution that may warrant further exploration, it is not so non-normal to invalidate the statistical procedures that were used. It appears that each variable was approximately normally distributed and thus combinations of those variables would also likely be normally distributed. Furthermore, the statistics describing each variable confirm the normality illustrated from the Q-Q plots where the ranges in the data

do not appear impeded by ceiling or floor effects and the variance within the data do not appear to be so significant as to inflate kurtosis, except for possibly that of SVT (Table 4.1).

Table 4.1:
Descriptive statistics of dependent and independent variables.

Variable	N	Minimum	Maximum	Mean	Std. Deviation	Skew ^a	Kurtosis
Dependent Variables							
GRADE	111	7	47	31.90	9.77	-.68	-.38
MCAS ELA	111	208	264	230.95	12.00	.34	-.52
Independent Variables							
ORF	111	46	211	125.68	34.54	-.07	-.13
RTF	111	17	127	68.41	22.91	.22	-.45
MZ	111	4	31	17.32	5.60	.19	-.45
WRT	111	2	42	21.58	8.75	.04	-.44
SVT	111	16	62	46.03	7.68	-.86	1.45

^aThe standard error for the skewness of each measure was .23.

^bThe standard error for the skewness of each measure was .46.

Interrater Agreement and Procedural Integrity

Fidelity in the administration and scoring of the individual assessments was checked on 20% of the sample for ORF and RTF. All examiners were highly trained using both measures, and have used the ORF measure frequently in their own professional experience. Regardless of their expertise with the measure ORF and their precise training of RTF, an Assessment Integrity Checklist along with synchronous testing was used to evaluate fidelity in the administration and scoring on one out of every five administrations. Using the criteria of two defined by the test authors, the agreement between the examiner and integrity assessor was computed using the percentage agreement formula, $(\text{Agreements}/(\text{Agreements} + \text{Disagreements})) \times 100$. Only 67% of the ORF integrity checks were within two points of one another. More striking, only 33% of the RTF integrity checks were within two points of one another. When correlations

were used to evaluate agreement, the examiner's scores correlated nearly 1.0 on the ORF measure and .97 on the RTF measure. An examination of the scatterplot (Figure 4.1) for each measure confirmed that the examiners reliably rank-ordered the students similarly, with more variability in the RTF score than in the ORF scores. With valid test administrations, highly trained and competent examiners and high correlations between examiners, all evidence suggests the two-point criterion for agreement between examiners is unreasonable to expect, and it is necessary to examine a more valid and realistic criterion for agreement.

Figure 4.1:
Scatterplots to examine inter-rater variability of ORF and RTF.

Additionally, interrater agreement was computed for the scoring of 22 of WRT tests. The raters were observed to agree .97 in their determination of the unique content words written for this sample of the test that were administered, indicating a strong degree of interrater reliability for the scoring of this procedure.

Examining Relationships between Variables

A correlation matrix was computed to compare the relationship between the dependent variable, GRADE, and each of the independent variables, the formative assessment measures, as well as the relationships between each of the independent variable. This correlational analysis revealed an informative picture of the function of

each assessment within the regression model as well as the contribution of each formative assessment measure to the construct of reading comprehension (Table 4.2). The relationship between overall reading proficiency, GRADE, and four of the formative assessment measures, ORF, SVT, WRT and MZ, was moderately strong (r ranged from .56 to .63); ORF, SVT and MZ each explained approximately 40% of the variance in performance on the GRADE while WRT explained approximately 31% of the variance.

There was a moderately strong relationship between ORF and each of the other formative measures of reading comprehension (r ranged from .49 to .72) with MZ and RTF exhibiting the strongest and weakest relationships, respectively. There was a moderate relationship between SVT and the other independent variables (r ranged from .47 to .57). Similar to ORF, the strongest relationship for SVT was with MZ and the weakest relationship was with RTF. There was a moderately strong relationship between MZ and the other measures with ORF exhibiting the largest coefficient ($r = .72$). These two measures include similar texts, although the tasks required for each measure differs. As with the other independent variables, MZ was most weakly correlated with RTF ($r = .47$). WRT was also moderately correlated with all of the IV's. WRT's strongest relationship was with MZ and ORF ($r = .59$ and $.56$, respectively) and its weakest relationship between SVT and RTF ($r = .49$ and $.47$, respectively). Of all the formative measures, RTF had the weakest relationship with GRADE ($r = .43$) and the weakest relationship with all the other formative measures. (r ranged from .45 to .49).

Interestingly, these data showed the teachers' perceptions of their students' reading comprehension abilities were better able to rank their students according to their GRADE performance than any other measure of reading comprehension ($r = .77$). Next

to GRADE, the teachers' judgements correlated most strongly with ORF ($r = .72$), but similarly well with all the other measures (r ranged from .61 to .69). Apparently, the teachers in these schools were better able to judge their students' reading abilities from their knowledge of their student's reading comprehension skills than were the test of comprehension. It is important to note assessment is a common practice in Reading First schools, and these teachers are well aware of their students' GRADE and ORF performance for as long as the student has attended the respective school.

In summary, the moderate relationships between performance on GRADE and the formative measures of reading comprehension indicate these independent variables may offer valuable information for predicting overall reading competence. The moderate relationships between each of the independent variables indicate while each measure may be a similar construct, they also measure a unique element of reading comprehension. It is unknown whether the measures are valuable instructional information, if teacher judgement is better at predicting reading ability, although the savvy of the sampled teachers in applying data to their instructional decisions must be taken into account when interpreting the transfer of these results to other teacher judgements

Table 4.2:
Correlation matrix for variables in the study.

	GRADE	ORF	SVT	MZ	WRT	RTF	TRank
Grade	1.00	0.65	0.59	0.67	0.57	0.46	0.77
ORF		1.00	0.57	0.72	0.56	0.49	0.72
SVT			1.00	0.57	0.49	0.47	0.63
MZ				1.00	0.59	0.47	0.69
WRT					1.00	0.45	0.61
RTF						1.00	0.53
TRank							1.00

Exploring Variability in Reading Comprehension

It was hypothesized that formative assessment measures of reading comprehension will account for the variability in reading proficiency beyond that accounted for by ORF. To test whether a relationship exists between GRADE and each of the formative measures, four multiple regressions equations were computed and the significance of the relationship between each measure and GRADE was examined while controlling for the variability in reading proficiency explained by ORF (Table 4.3).

Table 4.3:

Multiple regression analyses of ORF with measures of reading comprehension.

<i>R</i> -Squared/ (Change)	Predictor	Unstandardized Coefficient (SE)	<i>t</i> -value	<i>p</i> -value	Partial Correlation
$GRADE_i = \beta_0 + \beta_1 (ORF_i) + \beta_2 (MAZE_i) + \varepsilon_i$					
0.51/ (.08)	Constant	11.32 (4.09)	2.77		
	ORF	0.16 (0.04)	3.60	.000	
	MZ	1.18 (0.28)	4.30	.000	0.38
$GRADE_i = \beta_0 + \beta_1 (ORF_i) + \beta_2 (SVT_i) + \varepsilon_i$					
0.50/ (.07)	Constant	-5.62 (6.57)	-0.86		
	ORF	0.21 (0.04)	5.62	.000	
	SVT	0.67 (0.17)	3.90	.000	0.35
$GRADE_i = \beta_0 + \beta_1 (ORF_i) + \beta_2 (WRT_i) + \varepsilon_i$					
0.49/ (.06)	Constant	12.34 (4.14)	3.00		
	ORF	0.22 (0.04)	5.80	.000	
	WRT	0.55 (0.15)	3.68	.000	0.33
$GRADE_i = \beta_0 + \beta_1 (ORF_i) + \beta_2 (RTF_i) + \varepsilon_i$					
0.45/ (.03)	Constant	10.86 (4.52)	2.4		
	ORF	0.26 (0.04)	6.82	.000	
	RTF	0.13 (0.06)	2.31	.012	0.22

Using a one-tailed significance test, each model significantly predicted performance on the GRADE, and each formative assessment measure appears to contribute significantly to our understanding of reading proficiency above and beyond that explained by ORF (*p* ranged between .000 and .012).. Each model predicted a significant proportion of

variability in overall reading proficiency, where *R*-square ranged between .45 and .51. When controlling for the variability in ORF, the partial correlations ranged from .22 to .38. When considering the range of variability in ORF, an additional 3% to 8% of the variance in GRADE was accounted for by the reading comprehension measures.

Deriving a Model of Reading Ability

A multiple regression was computed to evaluate the significance of a comprehensive model of reading competence, where ORF, MZ, SVT, RTF, and WRT were used to predict reading competency as observed by performance on GRADE (Table 4.4). This model was found to predict significantly reading ability ($p = .000$) where the model accounted for more 57% of the variability in the dependent variable ($R^2 = .57$).

Table 4.4:

Overall regression model for reading proficiency

$$GRADE_i = \beta_0 + \beta_1 (ORF_i) + \beta_2 (MAZE_i) + \beta_3 (WRT_i) + \beta_4 (RTF_i) + \beta_5 (SVT_i) + \epsilon_i$$

Model	Sum of Squares	Df	Mean Squares	F	Sig.*	R-Squared
<i>Regression</i>	15521.67	5	3104.34	27.33	.000	.566
Residual	11926.02	105	113.58			
Total	27447.69	110				

Additionally, this multiple regression analysis was used to evaluate which of the predictors explained reading ability above and beyond ORF while controlling for the variability in scores accounted for by the other measures. It was hypothesized that some of the formative reading comprehension measures would be better predictors for evaluating reading proficiency than others. Using the Holm procedure, three reading comprehension measures were found to contribute significantly to the model of reading, specifically MZ ($p = .004$), SVT ($p = .010$) and WRT ($p = .024$) (Table 4.5) and one did not significantly contribute to the construct of reading as measured by GRADE, RTF ($p =$

.05). Thus three out of the four formative measures appear to be useful for predicting overall reading proficiency, excluding RTF.

Table 4.5:
Stepwise comparisons of independent variables

Stepwise comparison of independent variables						
Predictor	Unstandardized Coefficient	SE	t-value	Sig.*	$\alpha = .05/n^{**}$	
$GRADE_i = \beta_0 + \beta_1 (ORF_i) + \beta_1 (MAZE_i) + \beta_1 (WRT_i) + \beta_1 (RTF_i) + \beta_1 (SVT_i) + \varepsilon_i$						
Constant	-2.193	6.251	-3.51			
ORF	.101	.045	2.22	.029		
MZ	.766	.285	2.693	.004	.0125	Significant
SVT	.410	.174	2.363	.010	.0167	Significant
WRT	.303	.152	1.992	.0245	.025	Significant
RTF	.039	.054	.724	.2355	.05	

Predicting Academic Achievement

Using the most parsimonious model revealed in the sequential regression analysis, a logistic regression was computed to predict academic achievement as observed via student performance on the English Language Arts component of the Massachusetts Comprehensive Assessment System (MCAS) (Table 4.6). Unfortunately, the model of reading proficiency was not well suited for the prediction of pass or fail on the MCAS ($r = .23$) Where no independent variable in the model significantly predicted achievement on the MCAS.

Table 4.6:
Results of logistic regression analysis

Predictor	Slope	SE	p	$\alpha = .05/n^{**}$
$MCAS_i = \beta_0 + \beta_1 (ORF_i) + \beta_1 (MAZE_i) + \beta_2 (SVT_i) + \beta_3 (WRT_i) + \varepsilon_i$				
MZ	.14	.07	.04	.0125
ORF	.01	.01	.23	.0167
SVT	.05	.05	.28	.025
WRT	.004	.04	.91	.05
Constant	-8.25	2.21	.00	

Because a logistic regression restricts the range of variability in a measure by evaluating it dichotomously, it does not provide as much information to guide future research. As exploratory research, this final multiple regression analysis was computed to determine whether a model of reading, using both a measure of reading fluency and a measure of comprehension, would predict academic success based on state standards of grade level achievement. In fact, this model predicts academic success well, accounting for 66% of the variability in academic achievement as measured by the MCAS (Table 4.7).

Table 4.7:

Overall regression model for academic achievement

$$MCAS_i = \beta_0 + \beta_1 (ORF_i) + \beta_2 (MAZE_i) + \beta_3 (SVT_i) + \beta_4 (WRT_i) + \epsilon_i$$

Model	Sum of Squares	Df	Mean Squares	F	Sig.*	R-Squared
<i>Regression</i>	6870.13	4	1717.53	20.32	.000	.66
Residual	8960.65	106	84.53			
Total	15830.78	110				

Chapter 5

DISCUSSION

Summary of the Research

This research was conducted to help identify formative measures that may reliably and validly reflect general outcomes of reading comprehension. The purpose of this study was to examine the incremental and predictive validity of four formative measures of reading comprehension - MZ, RTF, WRT and SVT. Data on four independent variables, ORF, RTF MZ, SVT and WRT, and two dependent variables, GRADE and MCAS, were collected from 111 fourth grade students from two Western Massachusetts elementary schools. It was hypothesized that, due to the many subskills subsumed within the process of reading, formative assessment measures of reading comprehension will contribute more to our understanding of reading ability than simply the measure of oral reading fluency. Four multiple regression analyses were conducted to examine this hypothesis, and it was found that each measure significantly contributed to the understanding of reading proficiency as measured in GRADE beyond that of oral reading fluency alone. Additionally, a multiple regression analysis was computed that included ORF and all four reading comprehension measures to evaluate whether each contributed a unique element to the construct of reading. The measures MZ, SVT and WRT, in addition to ORF, were discovered to be significant factors in predict reading proficiency. Furthermore, it was hypothesized that valid formative measures of reading comprehension can be modeled in a meaningful way so as to predict performance on measures of overall academic success. While a logistic regression analysis, using ORF, MZ, SVT and WRT did not significantly predict the dichotomous result of academic

success as measured by the Massachusetts state assessment, a multiple regression analysis revealed this model predicts 66% of the variance on this dependent variable.

Results of the Analyses

Before it was possible to evaluate the validity of the formative reading comprehension measures, it was first necessary to examine the reliability of the scores for ORF, RTF and WRT. Reliability estimates for SVT and MZ have been well documented. SVT assessments that employ four passages with 64 test sentences have produced reliability estimates of .70 to .80 (Royer, 2001). Alternate form reliability for MZ has been observed between .62 and .93 (Guthrie, 1973) and test-retest reliability was reported at .83 (Manchester, et al., 2004). In this study, the correlation between raters for WRT was .97. Using the 2-point criterion published as acceptable interrater agreement by the publishers of RTF (Kaminski & Good, 2003), the interrater agreement for both RTF and ORF could be considered poor (.67 & .33, respectively). However, the high correlations observed between the scores of the raters suggest these two measure reliably rank-order students, but caution should be used when making decisions that require comparing the scores to some standard (e.g., cutscores).

The correlation matrix revealed an informative picture of the function of each assessment within the regression model. The moderately strong correlations between each of the independent variables and the GRADE (where r ranged from .56 to .63) suggest each measure explains an important part of the variance in reading proficiency in this fourth grade sample. The moderate relationships between each of the independent variables (r ranged from .45 to .72) indicate that while the measures evaluate part of the same construct in GRADE, they also explain a unique element as well.

Each measure was independently evaluated for its contribution in explaining the variability in GRADE above and beyond that predicted by ORF, and each measure was found to predict reading proficiency significantly beyond that of ORF alone. Each model that included ORF and one measure of reading comprehension predicted a significant and meaningful proportion of variability in overall reading proficiency (R -square ranged from .45 to .70). Furthermore, when controlling for ORF, the reading comprehension measures contributed an additional 3% to 7% of the variance in GRADE. Thus the measures of reading comprehension contribute to our understanding of reading proficiency for fourth-grade students above and beyond simply using a reading fluency score.

The formative reading comprehension measures were analyzed together in one model, controlling for ORF, to evaluate which could predict reading performance when considering the variance accounted for by all the others. Using the Holm Sequentially Rejective Multiple Test procedure, the three reading comprehension measures, MZ, SVT, and WRT, contributed significantly to the model of reading. RTF, however, did not contribute in explaining the variability in GRADE above and beyond the other measures in the model and ORF. Thus three out of the four formative measures appear to be useful for predicting overall reading proficiency. These three measures had enough power to predict GRADE performance despite controlling for the variability in GRADE accounted for by the other measures.

Interestingly, these measures may each explain a unique element of reading comprehension as explored in the literature on instruction and the cognitive theories supporting the instructional research. Three important theories around the development of reading comprehension have emerged. First, reading comprehension occurs as a result of

actively attending to the task of constructing meaning through strategic metacognition, and poor comprehenders can be taught to be aware of what they do not understand and correct the misunderstandings (Torgesen, 1977; Markman, 1979; Tarver, Hallahan, Kauffman, & Ball, 1976; & Wong, 1980). The MZ assessment can be considered a task of active awareness of meaning during the reading process. This task entails a series of obstructions in constructing meaning for which students must be aware of the block and seek out the correct meaning from the content of the passage, thus actively attending to the meaning as they read.

A second theory suggests students come to the task of reading with a schema with which they map on the new incoming content specific information and generate meaning from text as a constructive process where prior structural knowledge interacts with new information as meaning is constructed (Kintsch & van Dijk, 1978). The SVT assessment method was specifically designed to observe students' differential ability to construct meaning from a linguistic string within a text by its interaction with what the reader already knows (Royer & Cunningham, 1981). While some test sentences are taken directly from the text and others have the same meanings but do not use the same linguistic structure, students with strong comprehension skills will have an easier time judging the latter as having the same meaning as sentences in the text than poor comprehenders who may be mistaken by the lack of the linguistic connection. The SVT is arguably a good measure for evaluating the interactive process of meaning construction.

Finally, reading comprehension has been conceived as the development of a personal summary of a text that naturally occurs as a result of understanding (Kintsch &

van Dijk, 1978). Instruction has been designed around teaching students to perceive the main ideas and construct brief summaries about what they read. The WRT assessment may be a good tool for measuring students' ability to summarize what it is they remember from what they read. The WRT procedure does not depend upon an exact retell, but a summarization that includes important content words within the text and synonyms for words within the text.

For formative measures to be useful to teachers, it is important the measures reliably predict differences in academic achievement so as to identify which students are on target for success and which students are not so as to make instructional decisions to change the course for those low achieving students. Using the newly constructed model of reading proficiency that included ORF, MZ, SVT and WRT, a logistic regression was used to predict achievement on the Massachusetts's state assessment (MCAS). This model was unable to make a simple dichotomous prediction whether students would pass or fail the fourth grade state test. However, when a multiple regression analysis was conducted to capitalize on the full range of variability in the MCAS scores, the model of reading proficiency predicted 66% of the variability in MCAS.

Limitations of this Study

While these analyses produced interesting results that provide further insight into viable instrumentation for measuring reading comprehension, the results must be interpreted with caution. In this study, multiple analyses were performed using one data source and the same measures, thus, increasing the chance of capitalizing on peculiarities in the sample. All the results must be replicated using a new data source for each of the different analyses conducted in this study before exploring additional research questions

that arise from these results. In addition to inflating the Type I error rate due to the multiple analyses, one apparent irrelevancy in the testing environment should be considered when interpreting the statistical conclusions. Furthermore, there may be some variability in the scores of the tests due to testing fatigue of the students in the present sample. While these assessments did not require a lot of time from the students, testing for this study was conducted after the students had participated in a few weeks of testing, including all the end of the year assessments conducted within the school and the arduous state assessment conducted in every fourth grade. This random irrelevancy in the experimental setting may threaten the statistical validity of these results (Cook & Campbell, 1979).

There is one important threat to the external validity of the results of this study. The students in the sample were from schools that draw from low-socio-economic communities and had a history of poor performance on the state assessments. While both participating schools received grants to support their literacy programs and improve the outcomes for the students they serve, and as a result have made tremendous gains in their performances as measured by the state's assessment system, the large majority of students included in this sample were from less advantaged situations than may be represented in the general population. Therefore the results of this research may not be true of a more heterogeneous population of students.

Implications for Practice

At the point when decoding is automatized and proficient reading can no longer be inferred from reading rate alone, production-based assessment tools are needed that allow educators to observe and quantify indicators of reading comprehension

development. This research helped identify four measures of reading, ORF, MZ, SVT and WRT, that may be highly salient indicators for students who struggle in reading beyond the basic mastery of the code. It is clear that reading comprehension involves a complex confluence of many cognitive activities depending on the purpose of the reading task at hand. Such a complex construct necessitates a variety of tools that uniquely contribute to the variance within reading comprehension. Not only will the results of the combination of the tools provide an indication of overall reading comprehension ability, but an analysis of the individual measures could provide insight into a unique area of weakness that could guide instructional decisions.

In the last two decades there has been a shift in the fields of Special Education and School Psychology, and subsequently in general education, regarding how to decide which struggling learners will receive Special Education services. Historically, to receive special services the presence of a learning disability had to be identified. To do so, a school psychologist would typically investigate the presence of a cognitive deficit to explain the child's failure. Most often, this cognitive deficit was identified by observing a student's cognitive ability as measured by an IQ-score and deriving a predicted level of achievement. Then a standard achievement test was given to provide the actual performance score. The predicted achievement score is then compared to the actual achievement score to determine if a significant discrepancy exists between the student's predicted level of performance and their actual performance.

Since the inception of the IQ-Achievement Discrepancy Model for the identification of students who qualify for additional service delivery, this process has been scrutinized for its innumerable flaws. Some of these criticisms have serious

implications for struggling learners. For example, 70-80% of children identified as having a learning disability struggle in the area of reading and 90% of those identified have difficulty with word recognition. (Lerner, 1989). However, because reading behaviors as measured by standard tests of achievement observe only small samples of early reading behaviors and focus mainly on reading connected texts, and typical students only begin to read connected text in the middle of first grade, a significant discrepancy in reading can not be identified until a student is in third grade. By the time special educators are able to intervene, there is a minimal chance to change the trajectory of reading achievement for struggling readers compared to the chance interventionists would have had if the need for services was identified years earlier.

A second serious implication of using the IQ-Achievement Discrepancy Model is children with low-average and low IQ scores do not qualify for additional services because there is no statistical difference between their measured cognitive ability and their observed achievement. It is assumed that these children are performing as well as can be expected, and additional services would not help them achieve as children with more typical cognitive ability. However, there is no reason to assume children with low overall cognitive functioning can not learn to read and write just as well other children. In fact, evidence reveals there is no difference between the needs of slow learners and the needs of those identified as LD (Fuchs, Fuchs, Mathes, Lipsey & Roberts, 2001). Additionally, a Special Education selection process specifically designed to exclude some struggling learners from additional instruction aimed at closing the achievement gap goes directly against the No Child Left Behind Act of 2001 (NCLB).

Furthermore, the IQ-achievement discrepancy does not necessarily select students who are learning disabled, because it fails to take into account educational opportunities and the quality of students' past instruction. Until science-based curriculum was written into law, there was no means of standard control for determining which approach to literacy instruction should be chosen over another. A student who gets an average ability score on an IQ test and a significantly below average performance on an achievement test may not necessarily be learning disabled, but rather educational deprived as a result of a poor method of instruction or inferior teaching for one reason or another. That a child could be given a permanent label as learning disabled is a disturbing consequence of poor instruction and an important reason for the field of education to adhere to science-based practices as it evolves into a mature profession that truly reveres evidence (Carnine, 2000).

Because of these and many other documented flaws of the ability-achievement discrepancy model, there has been a shift in the last two decades from focusing efforts on inferring internal deficits in students to focusing on students' instructional needs and matching intensity of instruction to extensiveness of the need. Currently a response to intervention (RTI) model of disability determination is the most viable alternative to the discrepancy model classification system (Fuchs, Mock, Morgan & Young, 2003). The goal of an RTI model of service delivery is to offer a seamless system of academic support between general education and special education service so as to offer students a continuum of instructional support based on their needs and not some inferred inner-child deficits (Kovaleski, Tucker, & Stevens, 1996).

Most RTI models rely on the integration of science-based instructional practices, valid academic assessments to screen for levels of instructional need and provide a database of individual students' achievement information to make instructional decisions, along with a team-based problem-solving process. Most commonly, RTI has been implemented as a three-tiered model of instruction in the area of reading instruction (Fuchs, Mock, Morgan & Young, 2003; & Marston, 2005). In this prevention-based approach to instruction, all students are provided with a core instructional program that is evaluated for its inclusion of the five essential features of literacy instruction and supported by evidence to be effective for most children. All students in this level of instruction are assessed three times a year using formative measures of assessment to evaluate whether they are performing as expected and making sufficient gains so as to maintain grade-level performance. These screening procedures provide educators the opportunity to identify students who do not appear to be making adequate progress and modify instruction as necessary. For these students, a second tier of instruction is provided, whether it is more teacher-led practice in the scope and sequence of the core program, or something more strategic or different to close the achievement gap. These students are typically assessed one to two times a month to evaluate their progress as a function of the additional instruction. For students who do not improve their level of ability nor their rate of learning in a meaningful way with the second tier of instruction, a more intensive instructional opportunity is provided to them, including more time and more strategic instruction and their progress is closely monitored between two and four times a month. It is the students in this third tier of instruction who are often categorized as learning disabled because of the gap between what they are able to do and what is

expected of them, along with their slow rate of learning despite intensive and strategic evidence-based practice (Fuchs, 2003 & Burns & Senesac, 2005).

Formative assessment tools that provide a database of performance data, specifically using CBM strategies, play a critical role in the RTI model of instructional decision-making and disability classification. For RTI to work, valid and reliable measures are needed that allow educators to observe academic performance directly on a curriculum-based general outcome that indicates learning and meaningful academic gains. There are ample measures of early literacy skills that provide such data, such as measures of phonemic blending and segmenting, letter naming and letter sound fluency measures, phonics measures through nonsense word and real word decoding. For decades, ORF has been used as a valid and reliable measure of overall reading competence, providing a direct observation of both automatic decoding and reciprocally indicating reading comprehension. But these early literacy measures and ORF are less predictive of students' performance beyond third grade (Stage & Jacobson, 2001). Thus there is a need for reliable and valid formative measures of comprehension and higher-order reading skills so the RTI model can be equally useful for identifying the needs of older students as it is for younger students.

Directions for Future Research

The authors of the RAND (2002) reading comprehension study group recommend, "Any system of reading assessments should reflect the full array of important reading comprehension consequences (p. 54)." The cognitive activities involved in reading comprehension are diverse and complex, requiring purposeful reading, the integration of known text schema with new text specific information and the

creation of a personal summary. This study identified three formative measures of assessment that are aligned with processes theorized as essential to the broad array of activities involved in reading comprehension and thus reflect authentic outcomes. Research that logically follows this study is closely aligned with the research agenda delineated by the RAND study group. Studies that follow should answer the questions whether these three tools meet the minimal criteria for reading comprehension measures as defined by the RAND study group:

1. The capacity to reflect authentic outcomes;
2. Congruence between assessments and the processes involved in comprehension;
3. Developmental sensitivity;
4. The capacity to identify students as poor comprehenders;
5. The capacity to identify subtypes of poor comprehenders;
6. Instructional sensitivity;
7. Openness to intra-individual differences;
8. Usefulness in instructional decision-making;
9. Adaptability with respect to individual, social, linguistic and cultural variation; and
10. A foundation in theory and psychometrics (pp. 56-57.)

In addition to replicating the results of both the construct validity and the predictive validity analyses, further examination of MZ, SVT and WRT are necessary before determining whether they could be beneficial screening tools for students who have mastered basic literacy skills. It is hypothesized that MZ is a tool that allows us to observe active reading and clarification strategies; SVT measures the integration of new content to prior schema that generates new meaning; and WRT is a measure of a student's ability to express their personal summaries. Further research is necessary to explore the construct validity hypothesized as a result of this study.

Once validity has been examined, it is important to investigate whether these measures are sensitive to between-student differences and thus able to identify which

students struggle in reading comprehension. Additionally, with three formative measures available to assess different areas of reading comprehension, another valuable study is the investigation into whether it is possible to identify subtypes of poor comprehenders.

While research indicates MZ reflects student growth where ORF begins to wane in sensitivity (Jenkins & Jewell, 1993), another important research question would be to evaluate whether SVT and WRT are sensitive to typical student growth over time and how they reflect student learning as a result of instructional change. SVT as given in this study required thirty minutes to administer. An interesting study could explore the use of a shorter form of this assessment given multiple times over the course of a given time period. Royer & Cunningham (1981) report reliability estimates as low as .50 when less test passages were used. However, it would be interesting to see if using a series of one test passage with 16 test questions over multiple testing occasions would provide a reliable slope that could be used to evaluate student ability level and the rate of learning over time. A similar study to evaluate whether WRT could provide a reliable trajectory of learning would also be interesting. However before such a study could be designed, WRT should be examined more closely to establish the most reliable test administration and scoring procedures that are both simple to use in multiple forms. Such scoring procedures could include more specific content analyses based on meaning units.

In addition to within-student sensitivity to reflect the development of reading comprehension skills as a function of instruction, these tools must also be sensitive to differential abilities across students and thus inform instructional decisions within a three-tiered model of instructional practice. However, for such assessments to be useful as a measure of meaningful growth, they must truly reflect indicators of performance that

predict academic success. Not only should the analyses examining the predictive validity of ORF, MZ, SVT and WRT be replicated, but further investigations regarding how they predict future academic achievement are necessary.

The entire research agenda to explore effective instructional practices for reading comprehension depends upon the availability of reliable and valid tools to assess both ability level and rate of learning. Not only do group-level instructional decisions hinge on this line of research, but alternative models of disability identification also rely on the use of measures that reflect development and learning as a function of good instruction. For a three-tiered model of instruction to meet the needs of students effectively beyond third grade, assessment tools that reflect differential skill ability are essential. For an RTI model to identify accurately students who have mastered the phonics code, but still struggle to comprehend what they read, the development of such instrumentation to reflect skill ability, as well as differential rates of learning, are critical. This present study begins to provide a picture of an assessment system that has the capacity to reflect authentic outcomes of reading comprehension and are aligned with some of the important processes involved in comprehension. Innumerable studies emerge from here to begin to explore the answers to important questions necessary for the development of a meaningful assessment system for reading comprehension.

APPENDIX A

SUMMARY OF DEPENDENT VARIABLES FROM THE NRP STUDIES

Instructional Practice	Assessments	<u>Transfer</u>
Comprehension Monitoring	Detection of inconsistencies in logic or meaning/error detection Recall Question-answering Course achievement Standard comprehension tests	Near Transfer Near Transfer Near Transfer Far transfer Far transfer
Summarization	Recall Question-answering Multiple-choice questions	Near transfer Near transfer Near transfer
Question-answering	Recall Short Answer Test Look back in text observations	Near Transfer Near Transfer Near Transfer
Question-generation	Quality of questions generated Question-answering Standardized comprehension test	Near Transfer Near Transfer Far Transfers
Text Structure	Retell/Recall Short Answer tests	Near Transfer Near Transfer
Multiple Strategy Instruction/ Reciprocal Teaching	Observations of strategy use Content Area Achievement Standardized tests	Near Transfer Far Transfer Far Transfer
Graphic Organizers	Summaries Text Recall Standardized tests	Near Transfer Near Transfer Far Transfer

APPENDIX B

CONSENT FORM

Dear parent,

My name is Amanda Marcotte, and I am a student in the School Psychology Program at the University of Massachusetts in Amherst. I would like to conduct five simple and relatively unobtrusive reading and comprehension assessments with the fourth grade students at Stefanik School. I am requesting the participation of your child in my study.

The staff at Stefanik School use simple and efficient assessments to implement effective reading instruction to strategically meet the needs of every student in the school. The tools they currently have available to them focus primarily on pre-reading and early literacy skills for students in Kindergarten through grade three. The purpose of my study is to identify similar assessment tools focusing on reading comprehension that can be used with older students.

What I would like to do?

I would like to assess as many fourth grade students as possible in the variety of reading comprehension measures that show promise in helping teachers frame their instructional activities. There are a total of five assessments. Two of these assessments will be given individually to each student. First the students will be asked to read three passages for one minute and then retell what they remember about each passage for one minute. The total individual testing time for your child would be six minutes.

The remaining assessments will be administered to the entire class. These reading comprehension tasks will include summarizing a short story, identifying matching sentences and a passage with fill in the blanks. Total testing time for this group test will be approximately 40 minutes.

The final step of my research is to compare these five efficient assessment tools with data the district already has for your child. I would like access to their reading comprehension data as collected by the Stefanik School staff, and their MCAS English language arts score. I will use these scores to compare with my tools.

How will the information be used?

I am not interested in the results for any one child. I understand the private nature of your student's test scores and will protect the confidentiality of their GRADE and MCAS scores. All data will be compiled into one set and not analyzed for any individual child. While I plan to present the analysis of the data set to a professional audience and for professional publication, all individual scores will remain private.

Will your child or the classroom teacher benefit from participating in this research?

The results of these reading comprehension assessments may be very useful in planning instructional activities. Once the data are collected, I will share the assessment information with your child's classroom teacher and reading specialists. I will share with them the results of my reading assessments to give them an additional tool to continue providing effective instruction for all the students in their class.

Participation:

If for any reason you do not want your child to participate, please sign below and send this page to school with your child. If you choose to allow your child to participate, then you need not do anything more.

Sincerely,

Amanda M. Marcotte
 Doctoral Candidate
 University of Massachusetts, Amherst
 School Psychology Program

I do not want my child, _____, to participate in _____.

(Child's Name)

the reading assessments conducted by the UMass research team.

Parent/guardian signature

Date _____

APPENDIX C

ASSESSMENT INTEGRITY CHECKLIST

Page 10

DIBELS™ Oral Reading Fluency Assessment Integrity Checklist

Directions: As the observer, please observe setup and directions, time and score the test with the examiner, check examiner's accuracy in following procedures, and decide if examiner passes or needs more practice

Fine	Needs Practice	
✓ box to indicate Fine or Needs Practice		
☐	☐	1. Performs standardized directions verbatim: <i>Please read this out loud. If you get stuck, I will tell you the word so you can keep reading. When I say, "stop" I may ask you to tell me about what you read so do your best reading. Start here. Begin.</i>
☐	☐	2. Holds clipboard and stopwatch so child cannot see what (s)he records.
☐	☐	3. Starts stopwatch after child says the first word of the passage
☐	☐	4. For first word, waits 3 seconds for child to read the word. After 3 seconds, says the correct word, starts the stopwatch, and scores the first word as incorrect
☐	☐	5. For all words, if child hesitates or struggles with a word for 3 seconds, says the correct word and scores the word as incorrect
☐	☐	6. Puts a slash through words read incorrectly
☐	☐	7. Follows discontinue rule if child does not get any words correct in first five words
☐	☐	8. At the end of 1 minute, places a bracket (e.g.,]) after the last word provided and says "Stop."
☐	☐	9. Records the number of correct words.
☐	☐	10. Shadow score oral reading fluency with the examiner. Is he/she within 2 points on the final score?
☐	☐	11. Performs retell standardized directions verbatim: <i>Please tell me all about what you just read. Try to tell me everything you can. Begin.</i>
☐	☐	12. If the student does not say anything for 3 seconds, say "Try to tell me every thing you can." This prompt can be used only once.
☐	☐	13. If the student does not say anything or gets off track for 5 seconds, circle the total number of words in the student's retell and say, "Stop."
☐	☐	14. At the end of 1 minute, circle the total number of words in the student's retell and say, "Stop."
☐	☐	15. Shadow score the retell with the examiner. Is he/she within 2 points on the final score?

APPENDIX D

Q-Q PLOTS OF UNIVARIATE DISTRIBUTIONS

Q-Q Plots for Dependent Variables

Q-Q Plots for Independent Variables

Normal Q-Q Plot of ORF

Detrended Normal Q-Q Plot of ORF

Normal Q-Q Plot of RTF

Detrended Normal Q-Q Plot of RTF

Normal Q-Q Plot of Mz

Detrended Normal Q-Q Plot of Mz

Normal Q-Q Plot of WRT

Detrended Normal Q-Q Plot of WRT

Normal Q-Q Plot of SVT

Detrended Normal Q-Q Plot of SVT

REFERENCES

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

Ball, E. W. & Blachman, B. A. (1991) Does phoneme awareness training in Kindergarten make a difference in early word recognition and developmental spelling? *Reading Research Quarterly*, 26(1), 49-66.

Baumann, J.F., Jones, L.A., & Seifert-Kessell, N. (1993). Using think alouds to enhance children's comprehension monitoring abilities. *The Reading Teacher*, 47(3), 184-193.

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

Bean, T.W. & Steenwyk, F.L., (1984). The effect of three forms of summarization instruction on sixth graders' summary writing and comprehension. *Journal of Reading Behavior*, 16 (3) 297-306.

Beck, I.L. & McKeown, M., G. (2001). Text talk: Capturing the benefits of reading aloud experiences for young children. *The Reading Teacher*, 55(1), 10-20.

Brown, A. L. & Day, J. D. (1983). Macrorules for summarizing texts: The development of expertise. *Journal of verbal learning and verbal behavior*, 22, 1-14.

Brown, A.L., Day, J.D. & Jones, R.S. (1983). The development of plans for summarizing texts. *Child Development*, 54(4), 968-979.

Buck, J., & Torgesen, J. (2003) *The Relationship Between Performance on a Measure of Oral Reading Fluency and Performance on the Florida Comprehensive Assessment Test*. (FCRR Technical Report #1) Tallahassee, FL: Florida Center for Reading Research.

Burns, M.K. & Senesac, B.V. (2005). Comparison of dual discrepancy criteria to assess response to intervention. *Journal of School Psychology*, 43(5), 393-406.

Carnine, D. (2000). *Why education experts resist effective practices (and what it would take to make education more like medicine)*. Washington, DC: Thomas B. Fordham Institute. Available: <http://www.edexcellence.net/institute/publication/publication.cfm?id=46&pubsubid=698>

Carnine, D. W, Silbert, J., Kane'enui, E.J. & Tarver, S.G. (2004). *Direct Instruction Reading*, (4th ed.) Upper Saddle River, NJ: Pearson Education, Inc.

Chall, J.S. (1967). *Learning to read: The great debate*. New York, NY: McGraw Hill.

Cook, T.D. & Campbell, D.T. (1979). *Quasi-experimental design and analysis issues for field settings*. Boston, MA: Houghton-Mifflin Company.

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

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education*, 37(3), 184-192.

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

Dowhower, S. L. (1999). Supporting a strategic stance in the classroom: A comprehension framework for helping teachers help students to be strategic. *The Reading Teacher*, 52(7), 672-688.

Durkin, D. (1979). What classrooms observations reveal about reading comprehension instruction. *Reading Research Quarterly*, 14(4), 481-533.

Ehri, L.C. (2004). Teaching phonemic awareness and phonics: An explanation of the National Reading Panel meta-analyses.. In P. McCardle & V. Chhabra (Eds.), *The voice of evidence in reading research* (pp. 153-186). Baltimore, MD: Paul A. Brooks Publishing Co.

Englert, C.S. & Mariage, T.V. (1991). Making students partners in the comprehension process: Organizing the reading "POSSE". *Learning Disability Quarterly*, 14(2), 123-138.

Englert, C.S. & Mariage, T.V. (1990). Send for the POSSE: Structuring the comprehension dialogue. *Academic Therapy*, 25 (4), 473-487.

Fuchs, L.S. (2003). Assessing intervention responsiveness: Conceptual and technical issues. *Learning Disabilities: Research and Practice*, 18(3), 172-186.

Fuchs, L.S. & Deno, S.L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children*, 57(6), 15-24.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, 53(3), 199-208.

Fuchs, L. S., Fuchs, D., & Hamlett, C.L. (1989). Monitoring reading growth using student recalls: Effects of two teacher feedback systems. *Journal of Educational Research*, 83(2), 103-110.

Fuchs, L. S., Fuchs, D., Hamlett, C.L., Waltz, L., & Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review*, 22, 27-48.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical and historical analysis. *Scientific Studies of Reading*, 5(3), 239-257.

Fuchs, D., Fuchs, L.S., Mathes, P.G., Lipsey, M.W. & Roberts, H.P. (2001, August). *Is "learning disabilities" just a fancy term for low achievement? A meta-analysis of reading differences between low achievers with and without the label: Executive summary*. Paper presented at the Learning Disabilities Summit: Building a Foundation for the Future, Washington, DC.

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

Fuchs, D., Mock, D., Morgan, P. L. & Young, C. L. (2003). Responsiveness-to-Intervention: Definitions, Evidence, and Implications for the Learning Disabilities Construct. *Learning Disabilities: Research and Practice*, 18 (3), p.157-171.

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

Gersten, R., Fuchs, L.S., Williams, J.P. & Baker, S. (2001). Teaching reading comprehension strategies to students with learning disabilities: A review of the research. *Review of Educational Research*, 71 (2), 279-320.

Gersten, R., Woodward, J. & Darch, C. (1986). Direct Instruction: A research-based approach to curriculum design and teaching. *Exceptional Children*, 53(1) 17-31.

Glass, G.V. & Camilli, G.A. (1981). *"Follow through" evaluation*. Washington, DC: National Institute of Education.

Good, R. H., & Kaminski, R. A. (Eds.). (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.) Eugene, OR: Institute for the Development of Educational Achievement. Available: <http://dibels.uoregon.edu/>.

Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third grade high-stakes outcomes. *Scientific Studies of Reading*, 5(3), 257-289.

Grossen, B. (1995-96). Overview: The story behind Project Follow Through. *Effective School Practices*, 15(1), Available: <http://darkwing.uoregon.edu/-adiep/ft/grossen.htm>

Guthrie, J.T. (1973). Reading comprehension and syntactic responses in good and poor readers. *Journal of Educational Psychology*, 65(3) 294-299.

Guthrie, J. T., Siefert, M., Burnham, N. A., & Caplan, R. I. (1974). The maze technique to assess, monitor reading comprehension. *The Reading Teacher*, 28(2), 161-168.

Hare, V.C. & Borchardt, K.M. (1984). Direct instruction of summarization skills. *Reading Research Quarterly*, 20(1), 62-78.

Haring, N.G., Lovitt, T.C., Eaton, M.D., & Hansen, C.L. (1978). *The fourth R: Research in the classroom*. Columbus, OH: Charles E. Merrill Publishing Co.

Harris, T.L., & Hodges, R. E. (Eds.) (1995). *The literacy dictionary: The vocabulary of reading and writing*. Newark, DE: International Reading Association.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65-70.

Howell, K., & Nolet (2000). *Curriculum-based evaluation: Teaching and decision-making* (3rd ed.). Scarborough, Ontario: Wadsworth Thomson Learning.

Jenkins, J. R., Fuchs, L.S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology*, 95(4), 719-729.

Jenkins, J. R., & Jewell, M. (1993). Examining the validity of two measures of formative teaching: Reading aloud and maze. *Exceptional Children*, 59(5), 421-432.

Kaminski, R. A. & Good, R.H. (1989). Assessing early literacy skills in a problem-solving model: Dynamic indicators of basic early literacy skills. In M. Shinn (Ed.) *Advanced Applications of Curriculum Based Measurement*. (pp.113-142). New York: Guilford Publications.

Kame'enui, E. J., & Simmons, D. C. (2001). Introduction to this special issue: The DNA of Reading Fluency. *Scientific Studies of Reading*, 5(3), 203-211.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2) 163-182.

Kintsch, W. & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, 84(5), 363-394.

Klingner, J.K. & Vaughn, S. (1998). Using collaborative strategic reading. *Teaching Exceptional Children*, 32-37.

Klingner, J.K., Vaughn & Schumm, J. S. (1998). Collaborative strategic reading during social studies in heterogeneous fourth-grade classrooms. *The Elementary School Journal*, 99(1), 3-22.

Klinger, J.K., Vaughn, S., Dimino, J., Schumm, J. S., & Bryant, D. (2001). *From chunk to click: Collaborative strategic reading*. Longmont, CO: Sopris-West.

Kovaleski, J.F., Tucker, J.A. & Stevens, L.J. (1996). Bridging special and regular education: The Pennsylvania Initiative. *Educational Leadership*, 53(5), 44-47.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6, 293-323.

Lerner, L. (1989). Educational intervention in learning disabilities. *Journal of the American Academy of Child and Adolescent Psychiatry*, 28(3), 326-331.

Lysynchuk, L.M., Pressley, M., d'Ailly, H., Smith, M. & Cake, H. (1989). A methodological analysis of experimental studies of comprehension strategy instruction. *Reading Research Quarterly*, 24(4), 458-470.

Manchester, K., Marcotte, A., M., & Matthews. W. (2004). An initial reliability and concurrent validity investigation of the Dynamic Indicators of Basic Early Literacy Skills' Retell Fluency. Paper presented at the meeting of the National Association of School Psychologists, Dallas, TX.

Markman, E.M. (1979). Realizing that you don't understand: Elementary school children's awareness of inconsistencies. *Child Development*, 50(3), 643-655.

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

Marston, D. (2005). Tiers of intervention in responsiveness to intervention: Prevention outcomes and learning disabilities identification patterns. *Learning Disabilities*, 38(6), 539-544.

McKinney, J.D., Osborne, S.S. & Schulte, A.C. (1993). Academic consequences of learning disabilities: Longitudinal prediction of outcomes at eleven years of age. *Learning Disabilities Research and Practice*, 8, 19-27.

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

Moats, L. C. (1999) *Teaching reading is rocket science: What experts teachers of reading should know and be able to do*. Washington, DC: American Federation of Teachers. Also available on-line: <http://www.aft.org/pubs-reports/downloads/teachers/rocketsci.pdf>

National Assessment of Educational Progress (NAEP). (2005). *The nation's report card: NAEP 2004 trends in academic progress: Three decades of student performance in reading and mathematics*. (NCES Number 2005484). Washington, DC: U.S. Department of Education Institute of Educational Sciences. Available: <http://nces.ed.gov/nationsreportcard/>

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Bethesda, MD: National Institute of Child Health and Human Development. Available: <http://www.nationalreadingpanel.org/>.

Palinscar, A.S. & Brown, A.L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, 1(2), 117-175.

Parker, R., Hasbrouck, J. E., & Tindal, G. (1992). The maze as a classroom-based reading measure: Construction methods, reliability and validity. *The Journal of Special Education*, 26(2), 195-218.

Paris, S.G. & Jacobs, J.E. (1984). The benefits of informed instruction for children's reading awareness and comprehension skills. *Child Development*, 55, 2083-2093.

Paris, S. G., Lipson, M.Y., & Wixon, K.D. (1983). Becoming a strategic reader. *Contemporary Educational Psychology*, 8, 293-316.

Pressley, M. & Afflerach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Pressley, M., Brown, R., El-Dinary, P.B. & Afflerbach, P. (1995). The comprehension instruction that students need: Instruction fostering constructively responsive reading. *Learning Disabilities Research and Practice*, 10 (4), 215-224.

Pressley, M., Johnson, C.J., Symons, S., McGoldrick, J.A., Kurita, J.A. (1989). Strategies that improve children's memory and comprehension of text. *The Elementary School Journal*, 90(1), 3-32.

Reading Excellence Act, PL 105-277, 112 Stat. 2681-337, 2681-393, 20 U.C.S. §6661a *et seq.*

Reeve, J., Bolt, E. & Cai, Y. (1999). Autonomy-supportive teachers: How they teach and motivate students. *Journal of Educational Psychology*, 91(3), 537-548.

Rinehart, S.D., Stahl, S.A. & Erickson, L.G. (1986). Some effects of summarization training on reading and studying. *Reading Research Quarterly*, 21(4), 422-238.

Rosenshine, B. & Meister, C. (1994). Reciprocal teaching: A review of the research. *Review of Educational Research*, 64 (4), 479-530.

Royer, J. M. (2001). Developing reading and listening comprehension tests based on the sentence verification technique (SVT). *Journal of Adolescent and Adult Literacy*, (45)1, 30-42.

Royer, J. M. (in press). Uses for the Sentence Verification Technique for measuring language comprehension. *Progress in Education Research*.

Royer, J. M., & Carlo, M. S., (1991). Assessing language acquisition progress of Limited-English-Proficient students: Problems and a new alternative. *Applied Measurement in Education*, 4, 85-113.

Royer, J.M. & Cunningham, D.J. (1981). On the theory and measurement of reading comprehension. *Contemporary Educational Psychology*, 6 (3), 187-216.

Schank, R.C. (1972). Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology*, 3(4), 552-631.

Schumaker, J.B., Deshler, D.D., Alley, G.R., Warner, M.M. & Denton, P.H. (1982). Multipass: A learning strategy for improving reading comprehension. *Learning Disability Quarterly*, 5(3), 295-304.

Share, D.L., Jorm, A.F., Maclean, R & Matthews, R. (1984). Sources of individual differences in reading acquisition. *Journal of Educational Psychology*, 76 1309-1324.

Shaw, R., & Shaw, D. (2002) *DIBELS Oral Reading Fluency-Based Indicators of Third Grade Reading Skills for Colorado State Assessment Program (CSAP)*. (Technical Report) Eugene, OR: University of Oregon.

Shin, J., Deno, S.L., & Espin, C. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *The Journal of Special Education*, 34(3), 164-172.

Shinn, M. R., Good, R. H., Knutson, N., Tilly, W.D., & Collins, V. L. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review*, 21(3), 459-479.

Shinn, M. R., & Shinn, M. M. (2002). AIMSWeb Training Workbook: Administration and scoring of reading curriculum-based measurement (R-CBM) for use in general outcome measurement. [On-line]. Available: <http://www.edformation.com/>.

Singer, H. (1978). Active comprehension: From answering questions to asking questions. *The Reading Teacher*, 31, 901-908.

Singer, H. & Donlan, D. (1982). Active Comprehension: Problem-solving schema with question generation for comprehension of complex short stories. *Reading Research Quarterly*, 17(2), 166-186.

Snow, C.E. (2002). *Reading for Understanding: Toward a R & D Program in Reading Comprehension*. Santa Monica, CA: RAND.

Stage, S. A., & Jacobson, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review*, 30(3) 407-420.

Stanovich, P.J. & Stanovich, K.E. (2003). *Using research and reason in education: How teachers can use scientifically based research to make curricular and instructional decision*. Washington, DC: National Institute for Literacy. Available: www.nifl.gov/partnershipforreading

Stanovich, S.A. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, 21(4), 360-407.

Sweet, R. W. (2004). The big picture: Where we are nationally on the reading front and how we got there. In P. McCardle & V. Chhabra (Eds.), *The voice o evidence in reading research* (pp. 13-44). Baltimore, MD: Paul A. Brooks Publishing Co.

The White House. (1996, January 23). *State of the Union address of the President*. Available: <http://clinton2.nara.gov/WH/New/other/sotu.html>

The White House, (2001, September 8). *President emphasizes education reform in radio address*. Available: <http://www.ed.gov/news/pressreleases/2001/09/wh-010908.html>

Tarver, S.G., Hallahan, D.P., Kauffman, J.M., & Ball, D.W. (1976). Verbal rehearsal and selective attention in children with learning disabilities: A developmental lag. *Journal of Learning Disabilities*, 10, 27-35.

Torgesen, J.K. (1977). The role of nonspecific factors in the task performance of learning disabled children: A theoretical assessment. *Journal of Learning Disabilities*, 10(1), 27-34.

Torgesen, J.K. (1982). The learning disabled child as an inactive learner: Educational implications. *Topics in Learning and Learning Disabilities*, 2, 45-52.

Watkins, C.L. (1995-96). Follow Through: Why didn't we? *Effective School Practices*, 15(1). Available: <http://darkwing.uoregon.edu/~adiep/ft/watkins.htm>

Wendler, D. Samuels, S. J., Moore, V.K. (1989). Comprehension instruction of award winning teachers, teachers with master's degrees and other teachers. *Reading Research Quarterly*, 24, 382-400.

White, O. R., & Haring, N. G. (1980). *Exceptional Teaching* (2nd ed). Columbus, OH: Merrill.

Williams, K. T. (2004). *The Group Reading Assessment and Diagnostic Evaluation*. Circle Pine, MN: AGS Publishing.

Wong, B.Y.L. (1980). Activating the inactive learner: Use of questions/prompts to enhance comprehension and retention of implied information in learning disabled children. *Learning Disability Quarterly*, 3(1), 29-37.

Wong, B.Y.L & Jones, W. (1982). Increasing metacomprehension in learning disabled and normally achieving students through self-questioning training. *Learning Disabilities Quarterly*, 5(3), 228-240.

Wong, B.Y.L. & Wilson, M. (1984). Investigating awareness of and teaching passage organization in learning disabled children. *Journal of Learning Disabilities*, 17(8), 477-482.

6008-12