

UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

ARE OUR CHILDREN READING PROFICIENTLY AND HOW WOULD WE
KNOW? AN EXAMINATION OF STATE AND NATIONAL ELEMENTARY
READING ASSESSMENTS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

JULIE MARIE KALBFLEISCH COLLINS

Norman, Oklahoma

2007

UMI Number: 3271227

Copyright 2007 by
Collins, Julie Marie Kalbfleisch

All rights reserved.



UMI Microform 3271227

Copyright 2007 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

ARE OUR CHILDREN READING PROFICIENTLY AND HOW WOULD WE
KNOW? AN EXAMINATION OF STATE AND NATIONAL ELEMENTARY
READING ASSESSMENTS

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF
INSTRUCTIONAL LEADERSHIP AND ACADEMIC CURRICULUM

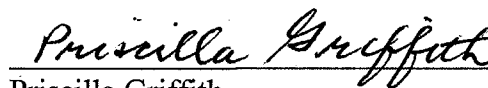
BY



Sara Ann Beach, Chair



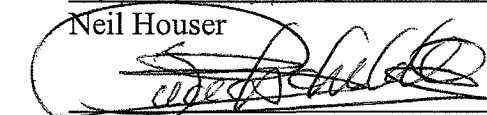
Gregg Garn



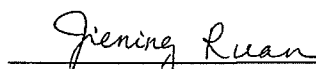
Priscilla Griffith



Neil Houser



Jos Raadschelders



Jiening Ruan

© Copyright by JULIE MARIE KALBFLEISCH COLLINS 2007
All Rights Reserved.

Acknowledgements

A person does not accomplish something of this magnitude alone. Many people have helped to encourage, inspire and assist me throughout the process of completing the requirement for this degree. I wish to express my appreciation to each of them, although these words seem quite incapable of expressing the depth of my gratitude.

First, I would like to thank the members of my committee, who have been patient, flexible, and supportive as they guided me through this process: Gregg Garn, Priscilla Griffith, Neil Houser, Jos Raadschelders, and Jiening Ruan. Most especially, the chair of my committee, Sally Beach, who has been a friend and mentor for many years, culminating in working with me through this process. Thank you for shaping the way that I look at the educational world and helping me to ask meaningful questions. I appreciate your patience and flexibility with me, meeting me after work hours, and helping me to spread my wings through this process. Sally, I appreciate your encouragement and belief in me!

I appreciate the help of others who assisted with statistical questions. Dr. Joseph Rogers, from the Department of Psychology, and fellow graduate student Clay Millsap, who spent time advising me about the use of the bootstrap procedure to answer my research questions. Many thanks go to Dr. David C. Howell, Professor Emeritus, University of Vermont, who advised and assisted me with his resampling software in order to complete the bootstrap computations. I am grateful for your patience and assistance in being able to complete these computations!

Thank you to members of my family who have believed in me, encouraged me, and patiently waited while I have been unavailable for many activities while I have been

working. Karen, thank you for your “long distance bops on the head” as I have worked to finish this endeavor this year. To Karen, George, Carl, members of your families, and Pat and Wallace, and all members of my extended family, thank you for always encouraging me and asking how the writing was coming along and for understanding my drive to take on this project. Mom, thank you to you and Dad, for your ongoing support and encouragement in all aspects of my life, especially while I have worked on this degree. Most especially, David and David, Jr., for the two of you have lived this more closely than anyone. Your support while I have followed this path through the many years it has taken me to complete it has meant the world to me! Through my ups and downs you have been there to listen to me and encourage me to finish!

Friends have also provided much needed encouragement through this process. Thank you Jennifer for providing unquestioned support since the day that I told you I was going back to graduate school. Thank you for always being there for me and for understanding the time constraints that it put on me! Thank you so much to the many colleagues at work and in graduate school who have been part of my support system over the years. I especially want to thank one of my first principals, Dr. Henry Walding, who encouraged me early in my career to return to graduate school, and two graduate school colleagues, Linda McElroy and Kris Akey, who have been such great supports to me through this process, offering suggestions and always being available in person and by telephone with a listening ear!

Finally, I dedicate this dissertation to my father. Dad, I wish you were here to talk to me about the statistics, attend the defense, and celebrate with me! Thank you for your inspiration, encouragement, and belief that I could achieve this milestone in my life!

Table of Contents

Abstract		x
Chapter 1	Introduction	1
	Definition of Terms	9
Chapter 2	Review of Related Literature	11
	Reading	11
	Factors Affecting Text Difficulty	15
	Factors Inherent in the Text	16
	Factors Dependent on the Reader	20
	Measuring Text Difficulty	22
	Assessment	27
	Reading Assessment	36
	Education Funding	47
	Summary	49
Chapter 3	Methodology	50
	Sample	50
	Sample Selection	51
	Data Sources	51
	Procedures	53
	Analysis	61
Chapter 4	Findings	64
	Variables	64
	Correlations	70

	Comparisons of state tests	75
	Comparisons of state tests and state and NAEP	81
	Summary	103
Chapter 5	Discussion	106
	Text Difficulty	107
	Passage Length	110
	Comprehension Levels	113
	Implications for Policy	117
	Limitations and Further Research	119
	Summary	122
	References	124
	Appendix A: Complete state information	134
	Appendix B: List of websites	137

List of Tables

Table 1	State sample grouped by per pupil spending	58
Table 2	State sample ranked by difference in percentage of student proficiency between state and NAEP scores	59
Table 3	Descriptive Statistics: State factors	64
Table 4	Descriptive Statistics: Test item factors	66
Table 5	Descriptive Statistics: Factors affecting passage difficulty	67
Table 6	Descriptive Statistics: NAEP comprehension categories	68
Table 7	Descriptive Statistics: Traditional comprehension categories	69
Table 8	Correlations of state variables	72
Table 9	Correlations of state variables, continued	73
Table 10	Correlations of state variables, continued	74
Table 11	Stepwise Regression Model 1	77
Table 12	Stepwise Regression Model 2	78
Table 13	Stepwise Regression Model 3	80
Table 14	Passage length bootstrapped comparisons	83
Table 15	Spache readability bootstrapped comparisons	84
Table 16	Powers, Sumner, Kearsley readability bootstrapped comparisons	85
Table 17	Constructed response test items bootstrapped comparisons	87
Table 18	Multiple choice test items bootstrapped comparisons	88
Table 19	General understanding test items bootstrapped comparisons	91
Table 20	Developing interpretation test items bootstrapped comparisons	92
Table 21	Making Reader/Text connections test items bootstrapped comparisons	94
Table 22	Content and structure test items bootstrapped comparisons	95

Table 23	Test items that can be answered without reading the passage Bootstrapped comparisons	96
Table 24	Literal understanding test items bootstrapped comparisons	98
Table 25	Inferential/interpretive test items bootstrapped comparisons	100
Table 26	Critical reading test items bootstrapped comparisons	101
Table 27	Application test items bootstrapped comparisons	102
Table A1	Complete state list	135
Table B1	Websites for departments of education and NAEP data	138

Abstract

The purpose of this study was to determine if differences exist between the NAEP and state fourth grade reading assessments. Specifically, the research questions focused on whether there were differences in text difficulty, text length, and depth of knowledge requirements between the state fourth grade reading tests as well as between the state tests and the NAEP. Sample text passages from 28 states as well as the NAEP were collected. The passages were analyzed for readability levels, text length, and whether or not introductions and illustrations were included with the text passage. The related test items were analyzed to determine whether they were multiple choice or constructed response and how they assessed comprehension by two scales, the four categories used by the NAEP, and five more traditional categories of comprehension.

The states were divided into groups by performance levels based on the difference in the proportion of students scoring at proficient levels and above between their state and the NAEP 2005 fourth grade reading assessments and into quartiles by states' per pupil spending. These groups were compared with each other and the NAEP to see if differences existed for any of the variables. Differences were identified between several groups of states and between the NAEP and state groups for several of the variables. Stepwise regression was used to determine if any of the independent variables predicted the difference in proficiency levels between the 2005 state fourth grade reading test and the NAEP. Three models were identified accounting for 34% to 43% of the variance of the difference in proficiency levels of the test scores.

Chapter 1

Introduction

Reading is one of the basic skills that students are expected to master in elementary school. Since the inception of public education in this country, reading has been a basic portion of the curriculum, as the 1647 Massachusetts Bay Colony Old Satan Deluder Act required public schools to teach children to read and write (Records of the Governor, 1647). From the goal of that original law, to enable children to read the Bible in order protect them from the devil, to current expectations that children read “on grade level” in order to advance from grade to grade, learning to read is one of the expectations of parents when they send their child to school. The measure of public elementary school students’ reading achievement across America has been a controversial topic for quite some time, spurred by policy makers and citizens looking for graduating students to be ready for the business world. Public outcry was heightened during the 1950s with the publication of *Why Johnny can’t read and what you can do about it* (Flesch, 1955). This book focused public attention on the whole-word method that schools were using to teach reading and helped to point parents toward phonics. The current focus is to have students reading “at grade level,” by meeting at least the proficient level on state assessments, by the end of 3rd grade, as required by federal mandates (Public law 107-110, 2002).

Currently, public schools are operating under increased scrutiny due to state and federal mandates aimed at holding schools accountable as measures of accountability have been increasing across the country for several years. Many states have implemented their own accountability systems. State accountability systems in thirty-two states use student achievement as an indicator of school success. Additional indicators used include

attendance and graduation rates (Education Commission of the States (ECS), 2006).

Based on these indicators of school accountability, states have a variety of systems in place to recognize schools and districts. Thirty-nine of the fifty states use at least one measure of reward or sanction with individual schools or districts. While rewards are in place in a number of states, sanctions based on achievement results are more common. Thirty-three states sanction schools and thirty states sanction districts based on student performance, while only twenty states reward districts and nine states reward schools based on performance (ECS, 2006).

Accountability is required under the auspices of the federal No Child Left Behind Act of 2001 (NCLB) (Public law 107-110, 2002) in the form of annual standardized testing administered by each state's department of education. NCLB is the most recent reauthorization of the Elementary and Secondary Education Act of 1965 (Public law 89-10, 1965). Under this federal legislation, reading and mathematics testing is required annually in grades 3-8, as well as once in high school. Science testing in at least one elementary, middle and high school grade is also required under NCLB beginning in the 2006-2007 school year (Public law 107-110, 2002). Some states also have additional testing requirements. The results of these tests determine the status of schools and entire districts as either making adequate yearly progress or being in need of improvement. The goal set by NCLB is for all students to read proficiently, or "on grade level," by the end of the 2013-2014 school year, as measured by state assessments (Public law 107-110, 2002). For most states, these requirements build on systems that were already in place in their state prior to the implementation of NCLB. State legislatures and departments of education have had to adjust accountability systems in their states in order to meet the

new requirements. One requirement is that states must use a single accountability system throughout the state with all public schools, using the same academic standards and the same assessments. In order to meet the mandates of the law, some states had to update their testing systems, adding tests at a state level or developing new assessments to meet the requirements. Another important change that has been implemented in some states has been developing tests to meet the requirements of aligning the tests with the state standards. In most cases this has included the development of criterion-referenced tests to replace the use of norm-referenced assessments that may have already been in place (Public law 107-110, 2002; ECS, 2006).

The requirements of the accountability systems have been questioned and publicly debated. The mandate that students and schools are judged on the outcome of one test has proven controversial. The tests are used for high stakes outcomes including whether or not a school is identified as being in school improvement status and, in some states, whether a student will be promoted or retained in their current grade. Using a single test for any of these determinations highlights the debate between the academic and political communities. The American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999), state in their standards that major decisions involving students should always be based on multiple measures. Currently NCLB is under consideration for re-authorization in Congress. One of the current hot topics in the debate is whether or not states should be required to participate in federal accountability mandates in order to receive federal education funding. The Academic Partnerships Lead Us to Success Act (A-PLUS), HR

1539, was introduced in Congress in March, 2007, and would allow states to take control of their own accountability efforts without losing federal funding (A-PLUS Act, 2007).

Tests take on a high stakes profile when the results are reported publicly (Popham, 2004; Guthrie, 2002; International Reading Association (IRA), 1999). While policymakers and the public would like accountability from their public schools, debate swirls around the best method for implementing it. It is thought that having tests that closely align to curriculum standards should allow educators, the public and policymakers to track the academic growth of the student population. However, score inflation can occur even with tests that are well aligned to standards (Koretz, 2005). The testing process should be set up to test a sample of a broad spectrum of knowledge, but the more familiar the tests and test formats become, the greater the chance becomes that teaching to the test could skew the results and cause the test to actually be testing a much smaller domain than intended (Koretz, 2005; Popham, 2004; Guthrie, 2002; IRA, 1999). Citizens across the country believe that the current emphasis on standardized testing will result in teachers teaching to the test rather than teaching a broad curriculum (Rose & Gallup, 2006), with parents of children in public school believing this at a higher rate (74%) than citizens as a whole (67%). Additionally, of those responding this way, the overwhelming majority believe that this result of teaching to the test is a "bad thing" (Rose & Gallup, 2006), with 72% of public school parents and 75% of citizens responding this way. However, citizens believe that testing is a necessary part of the educational system, with 58% of respondents selecting "not enough" or "about the right amount" of emphasis on achievement testing in the public schools, with 39% responding that there is "too much" emphasis on testing (Rose & Gallup, 2006, p. 46).

The testing debate continues to swirl around the use of the tests for accountability purposes versus instructional decision making purposes. In fact, high stakes tests are rarely used to inform instruction, although that is one of the reasons that proponents claim that they want to see them used. Due to the format of mass production and the time required to receive results for these high stakes tests, these tests are generally not used by classroom teachers to guide future instruction. Given the large number of standards to be addressed, the sampling process used to create these tests cannot include enough samples to assure mastery of individual objectives and standards (Popham, 2006)

In addition to the annual state exams, the National Assessment of Educational Progress (NAEP) is given across the nation. This test is used as a measuring stick of the nation as a whole and is often used to compare individual state results to the performance of students across the nation. The NAEP began in 1969 as a voluntary program to measure student achievement across the country. NAEP tests assess reading, mathematics, science, United States history, civics, economics, writing, geography, foreign language, and the arts at benchmark levels in grades 4, 8 and 12. The NAEP utilizes random sampling of students in selected schools and is used to report on state and national results by content and grade level. No student information leaves the building where the test is administered and as a result no student test results are reported or available. Results are disaggregated and reported as a whole population as well as for specific subgroups. While there may be a general feeling that students are not achieving as well as they have in the past, an examination of the NAEP scores over the years suggests that this is not true. For fourth grade students, the scale scores have remained amazingly stable since 1971, the first year that the reading test was administered. In

1971, the average reading proficiency score was 208 and in 1996 the score had increased only four points to 212 (McQuillan, 1998). The stability of the scores is seen considering that 10 scale score points on a NAEP test represents approximately one year's learning (Berth, 2006). The most recent results of the long-range NAEP data show that more progress has been made recently, with both President G. W. Bush and Secretary of Education Margaret Spellings citing fourth grade results that show more progress in the past five years than in the previous twenty-eight (U.S. Department of Education (USDE), 2007; Fuller & Wright, 2007). This statement refers to an increase from 212 to 217 on the NAEP long-range study data, although this increase was seen from 1999-2004, beginning well before NCLB implementation and leveling off in the latter years of that period (Fuller & Wright, 2007).

Under NCLB, participation in the 4th and 8th grade reading and mathematics assessments has become mandatory for states, as well as individual local districts, which chose to participate in federal education funding under NCLB (Public Law 107-110, 2002). Other content areas assessed by the NAEP continue to be voluntary in nature. The NAEP utilizes a sampling process in choosing students from across each state allowing the collection of state and national data to measure student success without creating high stakes situations for individual students, teachers, schools or districts since the information is reported only for states and the nation as a whole (IRA, 1999).

Although the NAEP is not the assessment used under NCLB to determine if states and individual school districts have met their accountability goals, it is often the assessment that policymakers and members of the media use to discuss the condition of public schools across America. A recent report by the United States Chamber of

Commerce grades states on their education systems (U.S. Chamber of Commerce, 2007). The report card includes nine categories for which grades were computed. One of the categories is "Truth in Advertising about Student Proficiency," in which grades were based on the difference between the percentage of students identified as proficient on state tests and the NAEP in 2005. In this report, only five states received the top score of "A" based on this score comparison (U.S. Chamber of Commerce, 2007).

NAEP scores are often used to discuss the success or failure of schools across the country, but often without any discussion of the similarities and differences with other assessments that are used. Assessment results based on the results of state tests often construct a much brighter picture of their schools' success than the NAEP, making it difficult for business leaders, parents, and citizens to hold the education community accountable (US Chamber of Commerce, 2007). A recent study has highlighted the fact that the proficiency levels identified by state tests and the NAEP are often not in agreement (National Center for Education Statistics (NCES), 2007). This study found that all but 10 of the 32 states giving a fourth grade elementary reading test in 2005 had cut scores for proficient scores falling below the cut-point for the NAEP basic cut score (NCES, 2007). This lack of alignment between the state tests and the NAEP may be fueling one of the proposals for the reauthorization of NCLB. In preparation for the reauthorization of NCLB, it has been proposed that states report their NAEP proficiency results along side their state test results annually (USDE, 2007).

The debate continues to swirl around the subject of school accountability and student assessment. Proponents of NCLB cite gains in the NAEP and state assessment scores as signals that NCLB is a policy that is positively affecting school change.

However, the same data can be interpreted in different ways, as exemplified by two recent reports. The US Department of Education (2007), in a review of NCLB results, has highlighted several states that are currently on track to achieve the mission of NCLB of having all students performing reading and math at grade level by 2014. These states are Delaware, Kansas, North Carolina, and Oklahoma. The attainment of this goal is based on student achievement on the states' own assessments. However, two of these same states were also in the lower tier of states in alignment to the NAEP assessment for fourth grade reading. Oklahoma and North Carolina were found to have state cut scores equivalent to NAEP scores of 182 (Oklahoma) and 183 (North Carolina) in comparison to the NAEP cut score of 208 for basic and 238 for proficient achievement (NCES, 2007). The other two states cited by the US Department of Education as being on track for meeting the achievement goals for mathematics and reading achievement were not included in the NCES study.

Considering the many critical issues concerning the assessment and accountability requirements in place across the United States, we need to examine the tests being used. While the results of these tests are reported and analyzed annually, the tests themselves have not undergone much scrutiny. If these tests are the one and only measure being used to judge whether students in American public schools are reading proficiently, it would benefit the system, the process, and the students for the assessments that are being used to measure this goal to be analyzed for similarities and differences. This analysis will help us determine what the tests are actually reporting to us and how we might be able to improve them.

This research project is designed with the goal of better understanding how we determine whether our students are reading proficiently, or “at grade level.” The goal of this study was to compare elementary reading tests used under state accountability systems under NCLB to one another and to the nationally administered reading test, the fourth grade reading NAEP. Specifically, the research addressed the following questions: Do significant differences exist between state passages and state and national passages regarding the difficulty of the passages? Do significant differences exist between state passages and state and national passages regarding passage length? Do significant differences exist between state assessments and state and national assessments concerning higher order thinking requirements of items compared in terms of depth of knowledge/higher order thinking requirements?

For the purposes of this study, the following definitions will apply to the use of these terms.

Definition of Terms

Reading proficiently: a designation earned by students on a selected state or NAEP fourth grade reading test representing “on grade level” or satisfactory performance on the given test.

Passage difficulty: variables affecting the potential difficulty of a text passage, including passage length, sentence length, vocabulary, whether or not an introduction to the text passage is included, and whether or not an illustration is included with the passage.

Higher order thinking requirements and Depth of knowledge requirements: both terms relate to the cognitive processing necessary for the student to successfully complete the test items related to the text passages. These terms refer to a range of processing from basic recall to relying on background information to understand the passage or to be able to make connections across passages.

Chapter Two

Review of Related Literature

This chapter will provide an overview of research related to the current study of fourth grade reading assessments. The literature review will include information about the process of reading and factors that affect the difficulty of that process for individuals as they learn to read. Research regarding assessment and related policy will also be included.

Reading

Simply stated, reading is the process of obtaining meaning from print; however, reading is a complex process involving the decoding of print from text in order to develop an understanding of what was read. Many factors contribute to the process of reading, including word identification, fluency, vocabulary, comprehension, knowledge of text, personal experience, social context and motivation (Snow, Burns, & Griffin, 1998; Pressley, 2006; Harrison, 2004). These factors work in conjunction with one another as children gain reading proficiency, although none of these factors have proven to be *the* key.

For this study, reading consists of both the processes of decoding and the construction of meaning. Each is necessary, but not sufficient, to encompass the process. Reading has not actually taken place unless readers have constructed a meaning based on their interaction with a text. The process of constructing meaning integrates the processes of decoding words on the page, attaching meaning to the words and phrases based upon the readers' own knowledge and experience, and drawing conclusions about the meaning of the text based upon the readers' interaction with the ideas that they have read.

Comprehension takes place when the reader's background knowledge interacts with the writer's purpose. An author sets out to convey a message to the readers, but readers construct their own meaning based not only on the text, but also on what they bring to the reading of the text as far as previous experience and motivation in the context of the reading. Readers may react differently to texts that they choose to read on their own as opposed to texts assigned to them for work or school assignments. The theory that brings these components together for many in the field of reading is the transactional theory of reading and writing (Rosenblatt, 1994). According to the transactional theory, reading and writing are seen as processes of constructing meaning, based on the context of the situation and the stance of the reader. Writers are the first readers of their work, constructing meaning as they construct the text. As with theories of social learning viewing people in constant negotiation with their environment (Berger & Luckmann, 1966; Rogoff, 1990), transactional theory places people in the same constant negotiation with literacy events, set in the context of their entire environment. The emphasis on meaning construction through the constant interplay of the reader with the cues in the text builds the metacognitive knowledge needed to interact with text and create meaning. This understanding means no text is ever understood in exactly the same way by two different individuals, because the understanding of the text resides with the readers based on all of the knowledge and experience that the readers already possess, which will affect their understanding of the writer's message. An important component of the transactional theory is the continuum of readers' stance, from efferent to aesthetic. Efferent reading is approached with the purpose of reading for facts while aesthetic

reading is done with the goal of living through the reading event by welcoming the feelings and perceptions that accompany reading for pleasure.

Skilled readers are able to read large portions of text with automaticity and fluency (Samuels, 2004). Automaticity is the process of being able to recognize words quickly and effortlessly. The theory of automaticity outlines the importance of internal attention to the process of cognition. Attention causes more issues for beginning readers than skilled readers since beginning readers spend more energy decoding text. As a result, beginning readers often have little attention remaining to attend to comprehending. The theory suggests that skilled readers have developed automaticity in word recognition, and as such, the decoding has become automatic and their attention can be used for comprehension. Automaticity is often developed from the wide reading that skilled readers have experienced, resulting in their contact with a large number of vocabulary words and variety of spelling patterns (Pressley, 2006; Harrison, 2004) and results in the reader being able to allocate more time and mental attention to comprehension than to decoding (Samuels, 2004; Alexander, 2005-2006). Fluency involves not only reading words at a good rate to ensure available mental ability to be able to focus on comprehension rather than interrupting the reading to decode unfamiliar words, but it also includes prosody, how smoothly a reader is able to read the text (Rasinski, 2003). High scores for fluency correlate to high comprehension scores. Skilled readers use more specific eye movements to aid the process of recognizing words, fixations on specific parts of text, saccades when the eyes are jumping to the next fixation, and regression, when the reader's eyes jump to a previous part of the text. It is also known that skilled readers read nearly every letter of the text, an important fact

because individual spelling differences are important in word identification. However, what the research has not yet been able to clarify is whether the difference in eye movements is a cause or a symptom of poor reading (Pressley, 2006).

The ultimate measure of skilled reading is comprehension. Within the realm of comprehension there are inferences that readers make automatically while reading, drawing on past experience to create general understandings or conclusions as they read. There are also more active comprehension processes requiring the reader to interact with the text during the reading process. In fact, skilled readers are involved in understanding text before, during, and after the actual reading of the text. Before reading, readers may preview the text format and familiarize themselves with the content by reading an excerpt or looking at illustrations. During reading, skilled readers adjust their reading based on their monitoring of their reading as well as their own purpose, interest and motivation, adjusting speed and rereading as necessary. Skilled readers also make active connections to their own previous knowledge or experience on the topic or related to the narrative. After reading, a skilled reader may continue to reflect on a text, rereading portions of it as needed, or referring back to notes that they may have made during the reading (Pressley, 2006).

An expectation and commonly held belief is that students spend kindergarten through second grade “learning to read” and then begin “reading to learn” in third grade. This speaks to the expectation that once students learn the basics of reading that they should be able to use reading as a tool to learn content. This may only be possible if students have been prepared to use reading for content learning in the primary grades (Duke, 2000). However, learning to read is not strictly a process to be accomplished in

the primary grades. A lifespan view of reading development has been proposed which includes six levels of reading competence that may last well past the primary grades. This continuum of reading competence provides more specific ways to describe a reader's proficiency without having to label students strictly by opposing views of "good" or "poor" readers. The six categories are: highly competent readers, seriously challenged readers, effortful processors, knowledge reliant readers, non-strategic processors and resistant readers (Alexander, 2005-2006). The two ends of the spectrum with this list are the highly competent readers and the seriously challenged readers, representing either end of the continuum, with the other levels representing steps along the process. Alexander (2005-2006) suggests that a view of reading as a lifelong developmental process would help the profession see reading instruction as a process to be continued through secondary education, rather than a process that many in the profession, as well as policy-makers and the public, expect to be completed by the end of elementary school.

For the purposes of this research study, the process of reading is viewed as involving both the processes of decoding and the construction of meaning. Both processes are important to the act of reading. Examining reading in this manner affects how the process of reading, and the assessment of reading, are approached for this study. Basing the process of reading partially on the construction of meaning and the stance that the reader takes toward the text affects the perspective of the text selections on the assessments and the difficulty factors that may be involved in the texts for the readers.

Factors affecting text difficulty

Students need a variety of strategies to be able to successfully navigate a variety of texts as they learn to read. Several characteristics of text affect the strategies that

students need to successfully comprehend the text. These characteristics fall under two broad categories: factors which are inherent in the text, such as length, cognitive density, sentence length, vocabulary, decodability, predictability, type of text and the inclusion of introductions and illustrations; and factors that are dependant on the reader, including motivation, interest, background experience, and setting a purpose for reading (Hiebert, 2002; Johnston, 1992; Chall, Bissex, Conard, & Harris-Sharples, 1994).

Factors inherent in the text

Text type. Traditionally, primary reading instruction has relied heavily on narrative texts. As a result, students often were not able to read proficiently as they made the transition from reading narrative text to reading nonfiction text (Duke, 2000). While there has been discussion of the need for more nonfiction text in reading instruction, a Duke (2000) study of twenty first grade classrooms found a mean of only 3.6 minutes per day of instruction using informational text, with some schools, especially those in lower socio-economic areas, having no instructional time during a school day with informational text. (Duke, 2000; Pappas, 2006; Pappas & Pettegrew, 1998). In addition to exposure to non-fiction texts, there are specific strategies that can help students comprehend nonfiction text (Pappas, 2006; Pressley, 2002; Duke, 2000; Guthrie & Mosenthal, 1986). The knowledge and experience required to read and understand narrative texts is different from the knowledge, experience and strategies that may be required to read and understand expository text (Allington & Cunningham, 2006; Pressley, 2002; Duke, 2000; Pappas & Pettegrew, 1998).

The type of text that readers are processing is important to the strategies that they use to approach the task. Genre is a critical feature of both spoken and written language,

engulfing the context of the exchange of information between speakers in a conversation or between author and reader with written text (Pappas & Pettegrew, 1998). Genre is often thought of as the difference between fiction and nonfiction text, but in fact there are many other differences between forms of text than those broad terms. Even within a particular genre the role of language plays an important part in how a reader may approach, decode and understand a text. Descriptive language, for example, may be used in a variety of types of texts within the broad classifications of both fiction and nonfiction (Pappas & Pettegrew, 1998). Text type may be broken down into more than just the two broad categories of fiction and non-fiction. Within the realm of fiction fall subcategories of realistic stories, fantasy, traditional tales, and poetry (Hoffman et al., 1994), as well as others, and within the realm of non-fiction may be traditional informational texts, but also some that weave together features from other types of text, such as information-narrative (Duke, 2000). The NAEP categorizes the texts used on the assessments into two broad categories, reading for literacy experience and reading for information (National Assessment Governing Board, 1994).

Reading fiction and nonfiction text differs not only in the approach that readers may take toward the process, but also in the strategies required to comprehend the text. Reading a narrative text generally requires reading the text from beginning to end and being familiar with story elements that readers will encounter, including setting, characters, and problem development and resolution. Nonfiction text is more likely to be read differently in real life situations and students should be taught strategies to assist them with this process (Duke, 2000; Guthrie & Mosenthal, 1986). Important features of reading non-fiction text include using features of the text such as tables of content, index,

heading, sub-headings, captions and glossaries to be able to locate information, use of text structures such as problem/solution, comparative/contrastive, and cause/effect, and the inclusion of graphical elements such as maps and diagrams (Pappas & Pettigrew, 1998; Duke, 2000).

A study of comprehension of expository text was conducted with 172 fifth grade students in Michigan (Wixson, 1984). The students were all identified as performing at average or above average levels on the reading comprehension portions of a standardized achievement test. The students were randomly assigned to either the control group or one of three leveled questioning groups. Each group read a text passage judged to be at the fourth grade level by Fry's readability formula that was between 165 and 175 words in length. Each passage was followed by five questions for the student to respond to in writing. Three of the questions were text explicit questions and two were text implicit questions. The questions were related to the ideas in the article, which had been previously ranked in importance. One week after responding to the questions the students were asked to write everything that they remembered about the passage. The titles of the passages were read to them to aid in their recall, but they did not reread the passage at the time of the writing. They were also asked to write about anything that they left out of their written response because they considered it to be unimportant.

Results of this study suggest that what students remember about expository text is related to the question that they are asked following the reading. While children seem to recall the most important parts of narrative text, this study suggests that whether teachers direct students to the important points or the trivial details of an expository text will determine what they remember about the subject.

Another study demonstrating differences in understanding types of texts was completed by Guthrie and Kirsch (1987). The participants in this study were adult employees of a manufacturing company, twenty electronics technicians and twenty-five electrical engineers. The study involved the participants reading journal articles, schematics, and manuals related to their job responsibilities and answering questions following the reading of each text. The participants also responded to a survey regarding their reading practices. The instrument had twenty-four questions that were designed to assess how much time the participants spent reading for different purposes-to gain knowledge, to find specific information, to keep abreast in the field, and to evaluate a document.

Factor analyses were conducted. No linear relationship of the four independent reading activities was found. The independence of the results suggests that reading comprehension and locating information in text are in fact separate factors in the reading process. This finding suggests that while proficient readers are those that can construct meaning from an author's message after decoding the text and having an interaction between their background knowledge and the new material, that proficient readers may need a totally different set of skills when their purpose is not to comprehend connected prose, but to find specific information, words or phrases in written material, whether it is prose or another type of text.

Another factor that may affect the difficulty of text is whether or not illustrations are included (Johnston, 1992). Illustrations in the form of drawings or photographs may help readers make connections between the text and their experience and assist with their understanding of a narrative passage. With content area reading, illustrations may take

the form of drawings, graphs, photographs, or charts that may help to explain the text.

Using these graphic features may assist the reader but may also require specific strategies by the reader (Pappas & Pettegrew, 1998). For example, a reader may need to make use of captions under photographs rather than only reading the main portion of the text.

Additional factors affecting the difficulty of texts include text format, cognitive density of the content, the level of reasoning involved to comprehend the text, decodability and predictability (Hiebert, 2002; Johnston, 1992; Chall, Bissex, Conard, & Harris-Sharples, 1994).

Factors dependent on the reader

Readability cannot be solely determined by quantitative formulas of the measurable factors in text passages (Zakaluk & Samuels, 1988), other factors also affect the difficulty levels of texts. One of the major factors affecting the difficulty of texts is the student's background knowledge or experience. Background knowledge may include experience reading a specific format or genre of text, or existing knowledge or experience with the content included in the passage. Preparing to read a passage is an important step in the process of connecting the current reading to a student's background experience or knowledge. In school this may be accomplished through conversation or a structured activity to connect the known to the new, such as a KWL chart, but with independent reading it may take the form of previewing a text or reading a book jacket. In a testing situation, this may be accomplished through the use of an introduction before a text passage to help prepare the student for the reading passage. A student's experience with the topic of the text can play an important role in whether or not the child understands the text.

Another factor that may affect the student's understanding is their ability to decode text and be able to have enough energy to focus on the comprehension of the text. Readers' skills come into play in helping a text be easier to read and understand than it is for those students who struggle with fluent reading and automatic decoding (Samuels, 2004).

Students' motivation may affect their ability to read a difficult passage if their interest is high enough. Struggling readers have been able to read texts above their typical reading level when they have high interest in the subject matter (Allington & Cunningham, 2006). Students' motivation may also be related to their previous experience with the topic or genre. However, this is difficult to standardize across a mass produced test as there is no way to assure that all passages will be equally motivating and accessible to all students.

Whether a reader is reading for enjoyment or for information affects the stance with which they approach the reading process. Reading aesthetically is reading for enjoyment, and is often commonly associated with reading fiction or narrative text. However, if students are being required to read a narrative text and interact with it in ways that they do not choose, they may not read aesthetically. Alternatively, efferent reading is reading done to obtain information. This is commonly associated with reading non-fiction texts, however, when readers read fiction texts for specific information to answer questions, they may also read efferently and bring the same strategies to the reading process (Rosenblatt, 1994).

Measuring Text Difficulty

Multiple factors of text may affect the difficulty of the reading process for individual readers. A variety of methods have been used over the years to attempt to measure these factors and assign a difficulty level to texts. This section will provide an overview of some of these methods.

Sentence length, structure and vocabulary, as measured by the number of words, the familiarity of the words, or the number of syllables included in the words, have been combined in a number of formulas. These quantitative formulas have attempted to measure the difficulty of texts based on a combination of these factors and have been commonly referred to as readability formulas. Readability may refer to one of three characteristics of text: legibility, interest in the writing, or how the style of the writing supports a reader's understanding (Klare, 1984). While the first characteristic is not an issue for mass produced texts for large audiences, and the second factor would be classified under factors dependent on the reader rather than the text, the third characteristic is the one that is generally under examination with readability formulas.

A typical way to describe difficulty of text in classrooms is with the percentage of words that students read correctly. If students can read more than 95% of the words correctly, the book is considered to be on an independent level, between 90% and 95% of the words correctly, the book is considered to be on an instructional level, and reading below 90% of the words correctly to be on the child's frustration level (Harris & Hodges, 1995; Fountas & Pinnell, 1996). Additionally, in order to assign these levels to these percentages of words read correctly, a student would also need to show appropriate levels of comprehension following the reading (Harris & Hodges, 1995). However, while this

type of readability system is useful for matching students to appropriate text in the classroom, it obviously will not work well for passages on tests since all of the text needs to be able to be read independently. One factor affecting readability identified by Johnston (1992) is teacher intervention, but this factor also cannot affect test situations since students must work independently.

Many readability formulas have been developed to quantify the effect of factors including content, vocabulary, sentence length, passage length, and sentence structure and determine appropriate levels for text passages. Generally, text difficulty levels are stated as grade level equivalents, but some formulas use other numerical results. A variety of mathematical formulas exist that seek to quantify the characteristics of text that may affect the difficulty. This is of interest not only to those in the field of education, but also to personnel in many other professional fields who want to be able to utilize literature at the correct readability levels for their audiences. While not all professionals agree upon the use of readability formulas, the formulas can be seen as an efficient method of analyzing the difficulty of texts by predicting the readability of existing texts. Well-known readability formulas compute their figures based on a combination of characteristics including number or percentage of difficult words, and number of words and/or syllables per sentence. Each formula defines difficult text differently. For example, the Flesch Reading Ease formula computes grade level by computing a formula that includes counting easy words, defined as those of one or two syllables, and hard words, defined as more than two syllables (Klare, 1984), while the Dale-Chall formula identifies difficult words by comparing the text to a list of 3000 common words expected to be known by fourth grade students (Dale & Chall, 1948). Words which do not appear

on the Dale-Chall list of common words are classified as difficult words. Some formulas are specifically designed to work with texts for specific ages or levels of readers. Two specific formulas designed mainly for younger readers are the Spache and the Powers, Sumner Kearl. The Spache readability formula is based upon the use of unfamiliar words and the length of the sentences in the text samples. The words are categorized as familiar or unfamiliar based on their inclusion on a list of common words for the Spache, based on readers in fourth grade and below. The Powers, Sumner, Kearl formula is computed based on sentence length and number of syllables per 100 words in the text sample. A more recently developed readability formula is the Lexile Framework for Reading (MetaMetrics, 2006). This formula reports a continuous range from 200L to 1700L based on measures of semantic and syntactic characteristics of the text. The text is measured in chunks of one hundred twenty five words and the Lexile level is computed based on the difficulty of the words and the length of the sentences.

It is important to note that readability formulas do not work well as production guides for writers, and the formulas are not recommended to be used in this manner (Klare, 1984). It is imperative that any readability formula be used as a starting point and not an absolute value. Each formula has its own starting point and computes the factors that it includes differently creating results that may not be directly comparative (Klare, 1984).

Another more recent method of examining text difficulty include an examination of texts on three measures of engagement, predictability, and decodability (Hoffman, et al, 1994). This system uses holistic measures of the content, sophistication of the language, and the design of the texts, however, the study included only first grade texts

from published textbook series from 1986, 1987, and 1993. The study found that the texts in 1993 were more engaging and more predictable than the earlier texts, but that the earlier texts were more easily decodable than the 1993 versions. There was no indication of how this analysis might fit within a scale of reading acquisition, and as such how it could apply to texts for older, more developed readers.

A popular way of expressing the difficulty level of books in elementary classrooms using the instructional strategy of guided reading is leveling books. This technique is a process used with the goal of matching books of appropriate difficulty with students who are ready for the challenges presented in the text. This process is usually done with books designed for emerging readers which may be small, eight page readers with one line of text per page, through picture books and early chapter books. The technique of leveling includes examining features such as the number of lines of print in the book, the size of the print, the space between the lines of print, the placement of the print on the page, the use of organizational features (i.e.: headings, table of contents), the use of illustrations, and the type of text (Fountas & Pinnell, 1999).

Another method of measuring the difficulty level of text that may also factor in the readers' background knowledge is measuring the inference load. This is accomplished by determining the event chain of a text, which connects the actions, events and states through a text. The text is divided into two types of clauses, tensed clauses that contain verbs inflected for tense, and untensed clauses containing noun and verb phrase complements. Tensed clauses are used to create the event chain. Each action within the text is identified as an action, a physical state, or a mental state. In a study by Kemper (1983), five individuals were trained in the process of identifying the event chains in texts

and they processed six text passages. Interrater reliability was determined by repeated measure ANOVAs. The five judges were found to agree on 88% of their inferences, with the best agreement shown for physical states at 93% and the worst agreement for inferred actions at 84%. This procedure was used to show that the process can be reliably executed.

For the study, two coders scored sixty-two passages covering a range of reading levels. Only the codes for the inferred actions, physical states and mental states agreed upon between the coders were used for further analysis. A regression analysis was used to determine the best fit of the variables to predict the inference load. The mean number of actions, mental states and physical states was totaled and divided by the number of words in the text passage to compute a density for each type of inference link included in the passage. The best fit of prediction of the inference load was to consider the stated mental states, stated physical states and inferred mental states. The analysis included separating the texts by narrative descriptions, scientific explanations, and historical accounts. The inference load analysis was found to have similar application across all three types of text. The inference load formula was found to be highly correlated with two popular basal series readability indexes for texts. This was found by comparing the inference load with sixteen text passages from the two basal series, resulting in correlations of .67 and .59. This technique was thought to better represent the view of reading as an interaction between a reader and a text and also provide not only a predicted difficulty level for a text, but also an explanation of why a reader may experience reading failure.

Assessment

The wide-spread use of standardized achievement tests across America began after the publication of *A nation at risk* (1983), which called for a national, although not a federally mandated, system of standardized tests to identify students for remedial and advanced learning opportunities and provide diagnostic information to teachers regarding student performance. The implementation of these tests across the country borrowed from the business world and offered rewards and sanctions to schools based on student achievement, contributing to the high stakes nature of these tests (Amrein & Berliner, 2002). In today's educational climate, sanctions are more prevalent than rewards (ECS, 2006; Amrein & Berliner, 2002), holding to the expectation that children and school personnel will improve performance through motivation to improve their status under the accountability system (Amrein & Berliner, 2002), with the emphasis primarily on student performance on tests in reading or language arts and mathematics. While testing may be conducted in other content areas, NCLB requires the use of scores from the reading and mathematics curriculum areas to determine the accountability status of schools and districts (PL 107-110, 2002).

One study of testing examined the use of high stakes tests and their effect on student learning. For the purposes of this study, Amrein & Berliner (2002), identified high school graduation tests as the high stakes tests to be examined. In states that require passing a test in order to graduate, high stakes tests for graduation were more likely to be implemented in states with large or quickly growing populations, with 76% of the states with the largest populations requiring a test and only 32% of smaller states requiring an exit examination. The tests were also more likely to be implemented in states with lower

per pupil spending amounts, with 60% of states in the bottom half of per pupil spending ranges to have a high school graduation test, and only 45% of states in the top half of per pupil spending ranges to require one. Additionally, high school graduation exams were found to be more common in states in the South and Southwestern regions of the country and those that had centralized governments. Of great concern to the authors was the fact that high school graduation exams were implemented more frequently in states with high populations of minority students, with 75% of states with higher percentages of African American students already implementing such tests, and more planning to do so by 2008, and only 13% of states with higher percentages of Caucasian students implementing tests, and plans for only 29% of them to be giving tests by 2008. The authors used a sample of 18 states that had high stakes graduation tests in place at the time of the research. In order to determine if results on the state exam actually reflected student learning or simply specific test preparation in the state, the research compared the results from the state exams to results in the state for college entrance examinations and the NAEP, although they admit that student motivation to do well on all of these tests may not be equal, and that the populations taking college entrance exams does not exactly match the population of students taking the high school exit tests. The authors argue that while short term improvement was seen on the SAT and ACT, long term improvement in scores or participation rate did not accompany the increases in achievement on the state scores, and as such conclude that the consequences attached to the high stakes assessments do not result in increased learning by students.

As part of this study, Amrein & Berliner (2002) also studied fourth and eighth grade NAEP data for the selected states. This was partially due to the fact that twelfth

grade NAEP data were not broken down by states, but that it could be assumed that states implementing assessment requirements at the high school level would also be implementing other school reforms which would apply to lower grades, and as such the NAEP results from fourth and eighth grades could still be an indicator of student learning. Their analysis of the NAEP fourth grade reading results for the 18 states in their sample showed that six states had positive gains on the NAEP between 1992 and 1998, ranging from 1 to 5 percentage increases in proficiency, four states had negative results, with scores decreasing from 1 to 5 percentage proficient in the same time period, and three states having neutral results with no increase or decrease in scores between 1992 and 1998. Five of the states did not participate in the NAEP during the time periods evaluated, since participation prior to NCLB was voluntary. The research also followed cohorts of students between the 1994 fourth grade and 1998 eighth grade NAEP reading assessments, to see whether the same group of students showed gains or losses. This analysis showed that nine of the states posted gains and four showed losses. During this time period, 69% of the states with high stakes tests showed improvement in their NAEP reading scores. Some of the other results that this study reported were attributed to fluctuating exclusion rates of English Language Learners and students with Individualized Education Plans, as 75% of the states implementing high stakes tests had higher than average exclusion rates in 1998, but the cohort results were not seen to be affected by this phenomenon and were seen as “real” gains, however, these gains in reading were noted to be the only positive real learning gains found in the study.

Rosenshine (2003) reanalyzed the data used by Amrein and Berliner (2002) and presented a different viewpoint with the results. Rosenshine used the same group of states

from the original research and also matched the sample with a control group of states that had not implemented high stakes testing. Rosenshine also took issue with the calculations and perceptions of increasing or decreasing scores used in the Amrein and Berliner study, because states were listed as increasing or decreasing in scores in comparison to the national averages, not on the exact data from their states. When comparing the results from the two groups of states, Rosenshine attributes a growth of 3.44 in the high stakes states during the same four year period on the fourth grade NAEP, but only a growth of 1.21 points in the states that did not implement high stakes testing. He states that the results were mixed among the states utilizing high stakes testing in each content area and grade level. He suggests that the high stakes attached to testing in those states may in fact be paying off with increased student learning and that the increased NAEP scores are probably not due to test preparation in the classroom or the accountability system in place in the state.

The original authors responded to the criticism and reanalysis by conducting a reanalysis of the data themselves (Amrein-Beardsley & Berliner, 2003). This time they used all available NAEP data for two sets of states, those with high-stakes tests in place and a control group that did not implement high-stakes testing. Their initial reanalysis showed that the control group had a significant increase in fourth grade reading scores from 1994-1998 of 2.1 points ($p < .05$) and that the experimental group of states using high-stakes testing had a significant increase of 4.3 ($p < .05$). However, they recalculated the analysis controlling for states that had increased exclusion rates on the NAEP. When this was figured in, the control group of states showed a significant increase in NAEP scores of 1.6 ($p < .05$), but the experimental group of states using high-stakes testing

showed an insignificant slight increase of 0.5 points on the NAEP. Their similar reanalysis on the fourth grade NAEP math test did show significant gains for the high-stakes testing states, but the fourth grade reading and eighth grade math NAEP results did not show that there was proof of increased student learning in the states implementing high-stakes testing.

Carnoy and Loeb (2002) conducted an analysis of whether external accountability systems increase student achievement. This study was conducted to assess the increasing use of accountability systems during the 1990s, prior to the increased federal accountability mandates of NCLB. To begin the study, the authors rated each of the fifty states on the degree of the external pressure on schools to improve standardized test scores. The ratings were based on state-selected criteria and ranged from zero for states with no statewide standards and testing at that time to five for states that test students from primary through secondary grades, sanction and reward schools, and require a high school exit test for graduation. The independent variables used for comparison for the study included the NAEP 1992 and 1994 reading scores for fourth grade white and black students, the resulting change in scores over that two year period, eighth grade mathematics 1996 and 2000 NAEP scores, per pupil revenue information, and population information for the states.

Findings from this study found that states with larger populations and higher proportions of minority students were more likely to have stronger accountability systems in place in their state. Additionally, states with lower achieving white students were more likely to implement strong accountability systems. A positive, significant relationship was found between the eighth grade math achievement gains across racial groups and the

strength of the state accountability system, suggesting the possibility that focusing on standards and high expectations on the assessments may in fact produce academic gains. The fourth grade gains were not associated as strongly with the accountability systems as the eighth grade gains were. However, the researchers did not find any relationship between the accountability systems and ninth grade retention, progression from eighth to twelfth grade or tenth to twelfth grade for white or black students, except for the possibility of a potential relationship between the accountability systems and ninth grade retention of Hispanic students.

A study by Marchant, Paulson & Shrunk (2006) examined the NAEP scores in relation to state characteristics, as previous studies had done, but they added a new dimension of adding demographic data to the variables. The researchers used regression to determine which variables might predict the NAEP scores. They divided the states into two groups based on high-stakes and non-high-stakes characteristics. When they compared the NAEP results based on the high-stakes characteristics they found that the characteristic of states with high-stakes environments predicted NAEP test scores when entered on their own, but when combined with demographic information the high-stakes environment was no longer a good predictor when family income and parent education levels were included. They warn that it can be misleading to look solely at states by whether or not they include high-stakes testing in their states. Additionally, the researchers suggest that it may not matter just whether a state has high-stakes testing requirements, but how long the state has implemented the high-stakes tests, as it is possible that the longer that the requirements have been in place the more effect it could have in predicting test scores.

A year long ethnographic study in two schools in New York was conducted to study the relationships of teaching, learning, and mandated testing (Mathison & Freeman, 2003). The researchers spent at least one day a week at each school. The majority of teachers in each school felt that the English Language Arts test in their state is a good test, focuses on higher order thinking skills, and includes important aspects of literacy including reading, listening, and writing; however, the teachers feel pulled in different directions, one to be the professionals that they feel that they are and the other to follow the mandates for the testing and prepare students appropriately by exposing them to the format, content standards, and scoring practices of the state tests. The teachers in this study not only gave the state mandated test, but they scored the test themselves at the school, which brought another layer of insight into the process. The teachers were trained with a video and worked in groups examining student papers and discussing specifics when they had questions about how to score specific answers. The authors suggest that the teachers feel that their professional judgment is being questioned as they are made to feel that their main job is to prepare students for the tests and that they are being judged only by whether students reach the preset benchmark score, not by whether growth is seen or whether students have gained ground in other areas, such as the affective domain, that do not register on the test.

The pressure to have students perform well on the achievement tests is great and is felt by teachers. "Teaching to the test" is a phrase that has mixed connotations. On the one hand, if tests are measuring state standards and content objectives, teaching to the test is seen as a teacher's responsibility and is seen as a "good thing," but teaching to the test more traditionally has referred to teaching test format and specifics for a test that would

help a student raise their score, possibly without knowing any more of the content that is being assessed on the test (Mathison & Freeman, 2003; Amrein & Berliner, 2002).

Another aspect of “teaching to the test” is that teachers may, consciously or unconsciously, live up to the adage that “what is evaluated is monitored.” The result could be that if multiple choice test items with only one correct answer are valued as a measurement of student achievement or teacher quality, teachers may teach more low level comprehension skills and provide fewer opportunities for students to read extended text and participate in personal, thoughtful responses (Graves, 2002).

The pressure associated with the performance of students on the tests used under the NCLB accountability systems is heightened by the fact that the tests are scrutinized, used as the factor affecting school and district accountability status, and are publicly reported, and is what characterized these tests as “high stakes” assessments (IRA, 1999; Tierney, 2000). One study examined high and low stakes tests used in states to determine if the pressure associated with high stakes tests caused students to not perform as well as possible and as such not accurately represent their learning. Low stakes tests were considered to be standardized tests administered but not used in the accountability process. This study found similar levels of performance by students on the tests in two states, with high correlations between the performance levels on the high and low stakes tests, suggesting that the “high stakes” nature of the tests does not affect student performance (Greene, 2003).

Professional reviews and organizations generally are in agreement with their recommendations about the use of test results for decisions about student and school progress. The American Educational Research Association, the American Psychological

Association, and the National Council on Measurement in Education recommend that no single test score be used as the sole data for a high stakes decision regarding a child, which would include promotion, retention and graduation decisions (1999). Similarly, the National Research Council's Committee on Appropriate Test Use (Heubert and Hauser, 1999) offers a similar recommendation that test score should be examined in context of a student's overall performance and that other data, such as teacher recommendations, grades and extenuating circumstances, should also be considered along with the test results when considering placement for a student.

The goal of using tests to measure student learning may be well intended, but data are conflicting in determining whether testing is benefiting students. A recent study confirms that results on state elementary reading tests are improving while the NAEP scores are not improving at the same rate (Fuller & Wright, 2007). This study examined state tests used under NCLB accountability systems and the long term NAEP data to try to determine whether NCLB accountability mandates were producing results. Long term NAEP data shows that fourth grade reading scores have improved between 1971 and 2004, with the largest portion of that growth coming from 1999 to 2004, starting prior to the signing of NCLB and coming in at the tail end of the 1990s states' efforts to improve standards and accountability. One of the goals of NCLB is to close the achievement gaps between racial groups. For fourth grade reading, scores for white students have increased about five points over a 13 year period from 1992 to 2005, while scores for Black and Latino subgroups fell initially from 1992 to 1994, but then increased over one and a half grade levels over the next 11 years. From 1994 to 2005, scores for fourth grade reading for Latino students went from 188 to 203, and for Black students from 185 to 200. These

figures suggest that the gap is closing on the NAEP, however, those scores remain 26 or more points below the scores for white students on the NAEP. The authors also cite two examples of state score reporting that have been followed since the implementation of NCLB. In trying to meet the requirements of NCLB, states must report proficiency levels of students, which are determined at the state level based on state standards and state tests. Many states already had assessments in place prior to NCLB, so it is possible to follow their progress prior to and since the implementation on NCLB. Texas has seen an increase in proficiency levels in comparison to the NAEP in their fourth grade reading scores since the NCLB requirements began, which was reported as a difference of 48% in 2002 and has grown to 53% in 2006, while Massachusetts has seen a drop in their score difference, from 7% in 2002 to only a 5% difference with the NAEP scores in 2006.

Another review of performance on NAEP and state assessments was conducted by Lee (2006). This review documented the long term trends as reported above by Fuller and Wright, but also provided some other statistical comparisons. Lee calculated discrepancies between the NAEP and state test results by computing ratios of the state proficiency rates to the NAEP proficiency rates. The ratios center around one, with results greater than one showing a relatively lower standard compared to the NAEP and results below one showing relatively stronger standards compared to the NAEP. The farther the ratios vary from the score of one, the greater the discrepancy. These ratios were computed across years that data were available for both NAEP and state tests at the same grade level and content area. The results for fourth grade reading ratios were all above the score of one, ranging from 1.28 to 5.01, suggesting that by these calculations all of the states had lower standards for reading proficiency than the NAEP. Within this

range the scores were clustered at the lower end. Only one state had a ratio above 5.0, no states had ratios between 4.0 and 4.99, and 3 states had ratios between 3.0 and 3.99. The majority of the states, 27, had discrepancy ratios between 2.0 and 2.99 and 15 of the states had ratios of 1.0 to 1.99. Three states did not have enough state data to compute discrepancy ratios for fourth grade reading.

Reading Assessment

Measuring students' abilities to perform the task of reading is challenging. Measuring all aspects of reading that are deemed important provides a unique challenge, and, as a result, some factors end up being overly emphasized due to their placement on assessments. This may happen during the assessment development process because portions that cannot be scored with reliability between scorers have to be deleted in order to assure test reliability. This process makes it more difficult to measure factors such as engagement and interpretation and places increased value on factors such as factual recall, vocabulary and reading speed (Tierney, 2000). Testing individual skills may result in students who show mastery on a number of individual skills but not on the process of reading as a whole (Nation of Readers, 1985). The use of high-stakes tests has not shown to be beneficial to reading achievement, takes valuable instructional time, and do not usually provide much specific information about a student's reading achievement (Afflerbach, 2005).

It is difficult to summarize reading skills in a single score. While the sheer number of standards implemented in some states can be overwhelming, it can be difficult to separate out the exact standard or objective being measured at one time and to assign an achievement level to a specific score that summarizes which skills and strategies a

student can perform if they score at a certain level. The complexity of literacy skills creates a situation in which scores cannot readily be combined (Tierney, 2000). Reading experts expect that proficient readers will be able to read selections with fluency and prosody and be able to discuss or retell the selection, but, this would require measuring this process with one-on-one interaction between teachers and students, and as such is not feasible for a large number of students. Since it would be difficult to standardize and implement reading and response sessions for large numbers of students across the country, mass produced tests must be used to measure reading skills with students working independently in order to meet the current testing expectations. As a result, reading tests typically include reading passages followed by questions about the passages to be answered after reading.

The differences required between reading non-fiction and fiction texts may be difficult to replicate in an assessment situation due to the testing format requiring text passages that may be shorter than what students read in school or everyday life. In the space available, the test passages may not be able to provide format and illustration support similar to what a student may experience when reading other non-fiction texts, such as textbooks. As a result, the demands placed on the student to perform successfully on a reading assessment may be affected by the type of texts included (National Assessment Governing Board (NAGB), 2004). The test difficulty may be affected by the choice of narrative, expository, poetry, or real-life reading passages, as well as whether or not the text passages are excerpts from literature that the student may have encountered previously or if the passages are written specifically for the assessment.

The NAEP reading test includes a variety of genres at the fourth grade level, grouped under two broad topics of “reading for literary experience” and “reading for information.” Reading for literary experience includes passages that enable students to explore literary language and events including plot, characters, themes and settings. These text passages may include selections from novels, short stories, poetry, plays, folktales, and biographies. Reading for information includes text passages that have information for the reader to learn about the world around them, including magazine articles, textbook selections, newspaper articles, speeches and essays (NAGB, 2004).

One previous study on the text difficulty of assessment passages was completed by Hiebert (2002). In this study she compared the text difficulty levels of two samples of norm referenced tests, two state reading tests, and two oral reading assessments. All of the samples represented third grade level reading on the respective tests. The tests were analyzed by three different text difficulty scales, critical word factor, Fry readability, and Lexile. The study found that all but one of the assessment text passages was measured in the second grade range of difficulty according to the Fry readability formula. Additionally, reading speed was seen to be an important factor within several of the tests as the norm-referenced tests were timed and one of the state tests had exceptionally long reading text passages (8 pages), requiring students to be able to sustain their comprehension over a lengthy text in order to answer the questions following the passage.

In order to assess the students’ understanding of the text, passages include questions following each passage. These test items involve challenges in themselves. The items need to be related to the passage preceding the questions. Just as the text passages

can be written at different levels of difficulty, items likewise can be on a continuum from easy to difficult. Readability formulas are designed to work with longer samples of text, usually 100 words or more at a minimum. As a result, it is generally inappropriate to measure the difficulty of test items using any of these formulas and paying attention to the match between the items and the related passages as well as the non-quantitative measures of readability (Oakland & Laine, 2004).

Homan, Hewitt, and Linder (1994) have suggested a method of determining the readability level of single sentence test items. Their formula is based on a stepwise regression model that determined which factors affected the difficulty level so that attempts could be made to better match students and levels of difficulty as related to test items. The formula was developed using sentences from comprehension sections of informal and standardized reading tests, which had been developed using a norming process rather than readability formulas to measure the difficulty of the texts. A total of 300 sentences were chosen from first through eighth grade tests and coded by grade level. From this sample, 180 of the sentences were randomly selected to be used in the regression model to find the variables that had the greatest effect on the difficulty level of a sentence. The model identified the number of difficult words, word length, measured as those with seven or more letters, and sentence complexity, represented as the average number of words per unit as factors affecting the readability level of the test items. When computing the readability level for individual test items, the formula is followed for each test item answer choice, and then an average for the item is computed.

Using this formula in a later study, Hewitt and Homan (2004) used the social studies and reading comprehension scores of more than 7,000 third grade students, more

than 7,000 fourth grade students, and nearly 7,000 fifth graders from a large urban school on the Comprehensive Tests of Basic Skills, Survey, Form A (CTBS) to analyze the difficulty of test items. The social studies scores and item analysis were used for the analysis. The reading comprehension scores were used to group the students into groups of high comprehenders (those with reading comprehension scores at or above the 50%ile) and low comprehenders (those with reading comprehension scores under the 50%ile).

The twenty items from the social studies subtest were examined for difficulty at the district and national level, looking at the percent of students answering each item correctly in the district and across the country. The difficulty levels for each item were computed with the Homan-Hewitt Readability Formula. The difficulty levels, as measured by the readability formula, were correlated to the percent of students answering each test item correctly. The results for each grade level showed negative correlations between the readability levels and the percentage of students answering correctly, suggesting that when the readability level increases for the test items, the percent of students answering the test item correctly decreases. The correlations were stronger at fourth (district = $-.72$, national = $-.62$) and fifth grade (district = $-.72$, national = $-.62$) levels than at the third grade level (district = $-.56$, national = $-.58$) (Hewitt & Homan, 2004). Further analysis showed an inverse relationship between the percentage of students answering the items correctly and the readability figures for the items, which was most pronounced at the fifth grade level. Combined with the correlation findings, the study suggests that the readability of test items may be an important factor in the ability of students to perform successfully on tests. The authors argue that with the current high stakes testing environment across the country that it is important that we look at this

important variable in test scores, although while this may be an important factor, I have not found evidence that this is a factor that test companies are examining in test development. Additionally, the authors argue that the formula should not be used to create test items at preset levels, but only to measure items that have already been thoughtfully written considering content.

Difficulty of test items does not only relate to the difficulty of the reading task, but the difficulty of the cognitive tasks required of the student. This includes the level of analysis required, from literal level comprehension to higher order thinking including summarizing, predicting, and analyzing. The difficulty levels of items on the higher end of the continuum may require students to compare and analyze different portions of a text or two different texts. Additionally, the difficulty may be affected by the reading passages on the tests that the items are associated with, which are often shorter than texts that children read in school and everyday real-life situations, resulting in fewer opportunities to develop characters and plot and involve the students in higher order thinking (Sternberg, 1991). The types of test items also affect the difficulty for the student. Test questions may be multiple choice, providing several answers for the student to choose from, or constructed response, requiring the student to write their own answer to the question. The choice of which type of items to include on the test may be driven by a variety of forces, including politics, finances, and/or concerns about implementation and reporting of scores. It may be most appropriate to report the test results separately for multiple choice and constructed response items, unless great care is taken to align the item stems and construction of the items (Rodriguez, 2003). The process of reading involves many aspects working in conjunction with each other in authentic situations, but

a reading test has the effect of removing the authenticity of the task and creating a simulated environment, which could affect the purpose of the reading task (Sternberg, 1991).

The NAEP asks questions related to each type of passage as they relate to four different aspects of reading. The first, forming a general understanding, involves understanding a text in a broad manner, and reacting to the text as a whole. Next, developing interpretation, involves reacting and responding to specific portions of the text and possibly making connections across portions of the text. Third, making reader/text connections, involves making real world connections by applying the text to personal real-world experiences. Finally, examining content and structure involves consideration of the content, form and organization of the text, including examining the author's purpose in writing the text and making comparisons between multiple texts (NAGB, 2004). The NAEP is constructed with at least half constructed response items and the remainder multiple choice test items.

Individual state departments of education and assessment companies may use other methods of classifying the comprehension levels or depth of knowledge levels of the test items. Questions on the tests are matched to a state standard or reading objective, and also generally classified by the level of cognition required, across a range. One method of doing this is described by Sadoski (2004) with comprehension levels ranging from literal, addressing recall of information stated directly in the text, inferential/interpretive level, requiring some level of interpretation across portions of the text, critical reading, involving the assessment or judgment of the content of the text, application, involving the construction of knowledge from the reading and the application

of that knowledge to other tasks or the application of related information to the understanding of the text, and appreciation, in which a reader is personally involved in the text.

A study was conducted with fourth, sixth and eighth grade students to assess whether students' knowledge of the type of question asked affected the answer they provided to the question (Raphael, Winograd, & Pearson, 1980). In this study, eighty students at each of the designated grade levels participated and were grouped by low ability and high ability based on the reading comprehension scores on a standardized test. The students read two passages and responded to eighteen comprehension questions, classified as text explicit, text implicit and script implicit based on how the reader would answer the question. Text explicit questions are those that have the answers stated explicitly in the text. Text implicit questions have answers that require the reader to inference across sentences or paragraphs in the text. Script implicit questions require the reader to draw on personal background experience in order to answer the question. The participants also categorized each of the questions into one of the three categories.

Results of this study suggest that the participants responded to the comprehension questions based on their prediction of what the question was asking them to do. The children provided more text based than knowledge based answers to the text-based questions and more knowledge based than text-based answers to the appropriate questions. The researchers also found significant effects for the students based on ability, showing that the high ability students were more likely to match the strategy with the type of question being addressed while the lower achieving students were more likely to

be unable to match the strategy to the type of question (Raphael, Winograd & Pearson, 1980).

A later study involved fifty-nine sixth grade students, grouped into three ability groups (high, average and low) based on teacher judgment, reading group membership and their reading comprehension score on a standardized achievement test (Raphael & Pearson, 1985). The participants were drawn from a larger pool of students that had already had severely reading disabled students removed. The students were randomly assigned to treatment and control groups within their ability groups. The students were trained on the types of questions, drawing on the same question types as the previous study. Text explicit questions were explained as "Right There," since the answers were directly from the text. Text implicit questions were explained as "Think and Search," because the student would have to think about the answer and look across sections (sentences or paragraphs) of the text to find the correct answer. Script implicit questions were defined as "On My Own," because the answers could not be found in the text but they would have to come from the student's knowledge base. Students in each group read two passages and responded to eighteen questions following each passage. All students responded to the same passage, and the second passage was adjusted for high and low ability readers. As students answered each question they also noted the type of question that they believed that they were answering. The treatment group received forty minutes of instruction each day for four days. The control group only participated in the assessment portion, with limited direction on the types of questions prior to completing the requested tasks.

Student responses were scored based on their correct categorization of the question, the quality of their response, being correct or incorrect based and being text-based or knowledge-based, and how well they matched their category to the type of answer they provided. Results suggest that training students did help them appropriately identify the type of question being asked, and ultimately match their process of answering it to the task. Students had the most trouble identifying text implicit questions by judging whether the information was in the text or in their knowledge base. The results also suggest that training students increased the quality of the answers that they provided, with both the average students and low level students performing at higher levels. The low level students made the greatest gains with answering text based questions following the training (Raphael & Pearson, 1985).

The transfer of the data from standardized tests has been a concern for professionals and citizens as well. One study of reading assessments looked at the congruence of the test results of a fourth grade reading test and teachers' assessment of student reading achievement. The study was completed in Ohio and included data from over 5,000 students in ninety-three districts across the state. Fourth grade teachers and principals of the students were surveyed about their assessment of the student's readiness for fifth grade, and the following spring the fifth grade teachers were surveyed as a follow up. The particular Ohio law mandating the use of these reading test scores is known as the Fourth Grade Reading Guarantee and the law states that the only way to overcome being retained as a result of failing the test is by recommendation of both the child's teacher and principal. For this reason, principals were also surveyed about the students' readiness to be promoted to fifth grade. The study found strong relationship

between the teacher and principal recommendations, although it was not certain whether the principal recommendations were based heavily on reliance on the teacher recommendations. The researchers found that the educators' opinions on whether or not a student would be successful in fifth grade matched quite well with the student's test score, determining whether or not the child could be promoted to fifth grade. They also found that teachers judged up to fifteen times more students that needed to be retained than actually were retained (1.2% of the fourth grade students). Another interesting finding was that teachers from districts across the state with varying degrees of success on the state test had opinions that matched the district success rate. Teachers from districts that performed at lower levels on the fourth grade test seemed to hold operational concepts of proficiency which were below those of the teachers from districts that performed at higher levels on the state test (Cizek, Trent, Cranell, Hirsh, & Keene, 2000).

Related factor of education funding

One of the many factors affecting public schools across the country is the amount of money spent on education in each state. Whether or not money matters in education is an often debated topic in the field of education. The process of education is expensive, and while additional funds may not solve all of the problems associated with education, many experts believe that additional funding would be beneficial. Education funding is traditionally a local responsibility, being funded by state and local tax revenues. Different states use different formulas to figure the funding for individual districts (Carey, 2003). Federal assistance for education began under President Lyndon Johnson's administration, with the bulk of the funding provided through the Elementary and Secondary Education Act of 1965 (Hacsi, 2002). This was the original legislation overseeing education

assistance from the federal government. The most recent version of this law is the No Child Left Behind Act. While the federal funding for education provided under this law is valuable to schools, it generally provides less than ten percent of funding for public schools across the country.

In an important meta-analysis on the subject, Hanushek (1989) conducted analysis on studies related to education funding. He reported that there was not research to support the use of smaller class sizes based on improved achievement from lower teacher/pupil ratios and he found no relationship between the quality of the facilities that improved funding can provide and student achievement. He did find support for increased teacher salaries and more experienced teachers, although he could not determine that these were the key variables in those classrooms in determining increased student achievement. In a re-analysis of the same studies, Hedges, Laine & Greenwald (1994) used different criteria to determine which cases to include and exclude, and came to different judgments about the importance of money in education. They argue that the reanalysis showed less of a relationship between some of the factors and student achievement, but argued that local authorities should maintain control of the decision making process for allocating resources. Hanushek (1994) responded that since teacher salaries are such a large part of the education budget, they are worth examining in terms of student achievement as well as numbers of students served in a class. While the debate continues today, attempts have been made to equalize the funding question by equating differing per pupil expenditure levels by considering the cost of living between locations (Carey, 2003).

Another analysis of education spending focused on the differences between funding schools with large populations of minority and special education students (Liu,

2006). He argues that the disparity of funding is increased by the method by which federal funding is allocated to the schools, creating larger gaps for the most at-risk students to make up. While funding issues often focus on whether money matters to schools, the debate should probably focus on the most efficient manner of using the allocations that are available (Hanushek, 1994). Perhaps “throwing money at schools” will not solve all of the problems in education, but using funds for purposes that show promise makes sense.

Summary

This study approached the process of reading as a process of decoding as well as meaning making. This approach laid the foundation for examining the reading assessments. This chapter has offered a review of literature related to the topics related to this study, including skilled reading, text difficulty, assessment, reading assessment, and the related factor of per pupil funding. This research review will serve as the basis for the decisions made in the study and the foundation for the search for answers brought about in some of the current research about why the reading test scores on state tests and the NAEP are not in better alignment (Fuller & Wright, 2007; NCES, 2007). While several studies have looked at characteristics of states and student demographics in relation to the comparison of scores (Amrein & Berliner, 2002; Carnoy & Loeb, 2002), none have examined characteristics of the assessments for evidence of why the scores may be different.

Chapter 3

Methodology

The goal of this research study is to create a detailed picture of how fourth grade elementary reading tests used in state accountability systems under No Child Left Behind compare to one another and how they compare to the nationally administered reading test, the NAEP. Specifically, the research addressed the following questions: Do significant differences exist between state passages and state and national passages regarding the difficulty of the passages? Do significant differences exist between state passages and state and national passages regarding passage length? Do significant differences exist between state assessments and state and national assessments concerning higher order thinking requirements of items compared in terms of depth of knowledge/higher order thinking requirements?

Sample

The sample for this study was a group of fourth grade reading assessments from twenty-eight states and the NAEP. Each state is required under the federal No Child Left Behind act to test students in third through eighth grade in mathematics and reading or language arts. Full implementation of this requirement was required to be implemented by the 2005-2006 school year (Public Law 107-110, 2002). Additionally, states may also test students in other content areas. The NAEP tests students in reading, mathematics, civics, United States history, and science in the fourth, eighth and twelfth grades. The fourth grade NAEP reading test was chosen as the sample of the elementary grade spans. As a result, fourth grade state tests were used as the basis for this study to facilitate comparison with the NAEP elementary reading assessment.

Sample Selection. The sample was chosen purposely by following several steps. Since the NAEP fourth grade reading test was last administered in 2005, the first step in selecting the sample of states was selecting those states that administered a state reading test to fourth grade students in 2005. Because this was one year before the required implementation of fourth grade reading tests under NCLB, not all states administered a fourth grade reading test that year. Next, state department websites from states that had administered a state fourth grade reading test in 2005 were examined to see if they had released test passages and items available for that test. If none were available online, the assessment division of the department of education was contacted about the availability of released test passages and items. If no released fourth grade passages and items were available, state department officials were asked about the availability and process of developing sample test passages and items. If the state department officials could verify that available sample test passages and items were representative of the test, the state was included in the sample. If no released or sample test passages or items were available, the state was removed from the sample. The result was a collection of released and sample fourth grade reading assessment passages and related test items from a sample of twenty-eight states.

Data sources. Data sources for the study were fourth grade reading test passages and the related test items from each of the selected states and the NAEP. Documents were collected electronically when possible, or in hard copy if electronic forms were not available. These documents were used to measure factors that affect the difficulty of the test for the students, including depth of knowledge requirements of the questions asked, passage length and difficulty of the text passages.

A test passage is any text or piece of text used in the assessment that has a group of questions connected to it. A passage may be a poem, short story, article, or real-life reading artifact such as a flyer or letter. The text passages may be selections taken from published literature, including magazines and novels, or may be selections commissioned specifically for the assessment. In some cases, a combination of texts is used with one set of items. States and the NAEP have different numbers and types of passages and items on the actual tests and released to the public.

The NAEP categorizes the texts on the test for fourth grade into two categories, “reading for literary experience” and “reading for information.” The first category, “reading for literary experience,” includes narrative texts that allow the reader to explore elements of story including events, characters, themes, settings, plots, actions and language. The texts in this category may include selections from novels, short stories, poems, plays, legends, biographies, myths, and folktales. The second category, “reading for information,” includes texts such as magazine and newspaper articles, textbooks, essays, and speeches that are read to learn information. In order to compare text selections on the state tests with the NAEP, these two categories, “reading for literary experience” and “reading for information,” were used to classify the passages used on the state tests.

Sample test passages and items are those that may be provided by the State Department of Education as examples of passages and items that could appear on the selected assessment. Released passages and items are passages and items that have appeared on an actual administration of the test and are no longer included in the test bank for future tests. These passages and items have been released to the public so that

teachers, students and citizens can examine the test format and content. The goal for the study was to use released passages and items for any tests for which they were available since those would provide the most accurate picture of that particular test. In some cases released passages and items were not available due to a state not releasing any passages and items to the public or the fact that none had been able to be released since the implementation of a new test for fourth grade reading in order to meet the assessment requirement of NCLB. In these cases, sample passages and items were included if the state department of education assessment personnel confirmed that the sample passages and items were developed under the same review process as the actual test passages and items and that they satisfactorily represented the potential format and content of the passages and items that may be included on the actual test. If the assessment department personnel from the specific state could not confirm that the sample passages and items represented the actual test, the state was removed from the study sample.

I had planned to collect two samples from each category of text from each selected state, to ensure characteristics of both categories of text would be represented in the analysis. This was not possible as some states did not have enough released passages and related items or appropriate sample passages and items for me to be able to collect two of each "reading for literacy experience" and "reading for information" text samples for the study. In selecting the text passages and items, I started by choosing from those released test passages and items from the 2005 test implementation, since that was the year of the most recent test scores from which the state selection was figured. If 2005 released items were not available, I selected from released passages and related items or sample passages and related items and worked through the past five years of available

passages and items. If more than one of each type of passage were available, I randomly chose one of the “reading for literary experience” and one of the “reading for information” text passages for inclusion in the study. Additionally, released passages and items from the 2003 and 2005 fourth grade reading NAEP assessments were collected from the NAEP website. Selections were collected from a total of twenty-eight states tests. In all, twenty-six “reading for literary experience” selections were included and twenty-five “reading for information” selections were included as a result of a few states not having both categories available. Additionally, four NAEP passages were collected along with the related test items. Two “reading for literary experience” and two “reading for information” passages were collected from the NAEP. Copyright restrictions prevent including the passages and test items.

Procedures. The following steps were followed to collect the data for this study. Step 1: The sample of twenty-eight states was selected. This was accomplished by selecting states that administered a fourth grade reading test in 2005. I searched the state department of education website for each state and for the NAEP. Within each website, I looked for the portion addressing student assessment and searched for released fourth grade reading test passages and items for the study. I created a spreadsheet to keep track of each state and the documents needed for the study. If any or all of the necessary documents were not available via the state website or the NAEP website, I found the contact information for the appropriate personnel in the assessment division on the website and inquired about the items needed or the open records process for that particular agency. I followed up on the return electronic mail messages and made contact by telephone when necessary. If no released passages and items were found, the

assessment department was contacted to find out if any released passages and items were available. If there were no released passages and items available, then the availability of sample passages and items was checked. If sample fourth grade passages and items were available, personnel from the assessment department of the state's department of education were asked to verify that the sample passages and items were developed through the same, or a similar, process to the passages and items used on the actual tests. If it could be verified that the sample passages and items were representative of the actual test, the state was included in the sample. If it could not be verified that the samples were representative of the actual test, the state was removed from the sample. Appendix A contains the complete list of states and their characteristics used during the sample selection. Appendix B contains a list of the state websites from which the released and sample test passages and items were obtained.

Step 2: The states in the sample were grouped for comparison. The twenty-eight sample states were grouped into quartiles based on per pupil funding for the states and into achievement groups based on levels of difference in proficiency on the NAEP and the state fourth grade tests.

Step 3: I downloaded the selected documents from the appropriate state and NAEP websites. Whenever possible, electronic forms of the documents were collected to facilitate computer use to analyze the documents. If test passages were only available in hard copy the selected passages were typed and saved on computer to be used for electronic readability and length computations.

Step 4: I created a database of the sample and released test passages and items. The database summarized the features regarding each test passage and the related test items, including genre, passage length, readability, and cognitive requirements of the test items.

Analysis. Once the sample was identified, the states in the sample were divided into groups for comparison in two ways. One way was based on per pupil funding criteria and the other method was based on the difference in percentages of students reading at proficiency level or above between the 2005 state and NAEP fourth grade reading tests. While it is difficult to isolate education funding as a factor in student achievement, research has suggested that wealth and expenditure levels, while not solving schools' problems alone, may be connected to increased student performance (Hacsi, 2002; Hedges, Laine & Greenwald, 1994). Due to the potential relationship between education funding and student achievement, the first step in grouping the state sample was based on per pupil expenditures across states. The state sample was ranked based on the per pupil expenditures for the 2003-2004 school year (US Census Bureau, 2006), since this was the most recent data available, and divided into quartiles. This information can be seen in Table 1.

States were also grouped by how their fourth grade students performed on the 2005 state reading assessment as compared with the 2005 NAEP elementary reading assessment results. Each state department of education assessment department has a designation for satisfactory performance on their elementary reading performance. The achievement level may be referred to as proficient, satisfactory, or at grade level. The percentage of students in each selected state scoring at this designated level on their specific state fourth grade reading assessment was compared to the percentage of students

from that state that performed at the proficient level or above on the NAEP. The difference between these two scores was computed for each selected state. The states were ranked based on the difference in the proportion of students scoring proficient or above. The list was divided into four groups based on these levels of proficiency differences: states with less than a 25% difference, states with differences between 26-40%, states with differences between 41-49%, and states with differences of more than 50% in proficiency between their state test and the NAEP. These groups are illustrated in Table 2. All of the state proficiency percentages were higher than the NAEP proficiency percentages.

The text passages were first categorized according to the two NAEP categories of “reading for literary experience” and “reading for information.” Passage length was computed for each passage and noted by the number of words included in the passage, including words in an introduction to the passage, if a related introduction was included.

The readability of each passage was determined based upon two formulas. Since the processes for determining readability are different for each possible readability formula, the results cannot be averaged together, but using multiple formulas allowed for comparison. The Spache and the Powers, Sumner, Kearsley were the readability formulas selected. Both the Spache and the Powers, Sumner, Kearsley are used to determine readability levels for students in kindergarten through seventh grade. The Spache formula is based on the length of the sentences in the text and the use of unfamiliar words. The Power, Sumner, Kearsley formula is based on the number of syllables and length of sentences in the text passage. The readability for each passage was computed electronically using Readability Studio software from Oleander Solutions. For purposes

Table 1

State sample grouped by per pupil spending

<u>Quartile</u>	<u>State</u>	<u>Per Pupil Spending</u>
1	New Jersey	\$12,981
1	New York	\$12,930
1	Connecticut	\$10,788
1	Massachusetts	\$10,693
1	Maine	\$9,534
1	Wyoming	\$9,363
1	Wisconsin	\$9,226
2	Maryland	\$9,212
2	Michigan	\$9,072
2	Ohio	\$8,963
2	New Hampshire	\$8,860
2	West Virginia	\$8,475
2	Montana	\$7,763
2	California	\$7,748
3	Georgia	\$7,733
3	New Mexico	\$7,331
3	Washington	\$7,243
3	Louisiana	\$7,209
3	South Carolina	\$7,184
3	Texas	\$7,104
3	Kentucky	\$6,888
4	Florida	\$6,784
4	Arkansas	\$6,740
4	North Carolina	\$6,702
4	Tennessee	\$6,504
4	Mississippi	\$6,237
4	Oklahoma	\$6,176
4	Arizona	\$6,036

Table 2

State sample ranked by difference in percentage of student proficiency

Performance Level	State	'05 state %age above NAEP
4	Georgia	61
4	Tennessee	60
4	Oklahoma	58
4	West Virginia	55
4	North Carolina	53
4	Maryland	50
4	Michigan	50
4	Texas	50
3	Wisconsin	49
3	Mississippi	48
3	New Jersey	45
3	Washington	44
3	Louisiana	44
3	Arizona	44
3	Ohio	43
3	Florida	41
2	Montana	39
2	New York	37
2	Kentucky	37
2	New Mexico	32
2	New Hampshire	30
2	Connecticut	29
2	California	26
1	Arkansas	22
1	Maine	18
1	Wyoming	13
1	South Carolina	10
1	Massachusetts	6

of comparison through this study, the same formulas were used for all passages, even though these formulas are not currently used by NAEP and may not have been used by the identified state being studied.

Two other factors affecting the potential difficulty of the text passages were recorded. Each passage was examined for the inclusion of an introduction to the story as well as for whether one or more illustrations were included. The introduction was counted if it helped to prepare students for the story and did not include only directions to read the passage and answer the questions. An illustration was noted if at least one of any type of illustration, drawing, photograph or graph was included with the text passage.

Test items were first identified as multiple choice or constructed response. The related items were then categorized for higher order thinking requirements by two methods. First, the items were classified in the same manner as those used on the NAEP to facilitate comparison between the state tests and the NAEP. The NAEP uses four aspects of reading to classify the assessment questions: forming a general understanding, developing interpretation, making reader-text connections, and examining content and structure. While classifying test items according to these four NAEP aspects of reading, a fifth category had to be added to categorize questions from the state assessments that did not fit into any of the four NAEP categories. The items were also analyzed in a more traditional manner, since this may more closely mirror how many states approach the task, by identifying the level of comprehension of the item: literal, inferential/interpretive, critical, applied, or appreciation (Sadoski, 2004). I computed a proportion of each type of question for each individual passage included in the data sample being studied. Additionally, a category was added to classify questions that could

be answered without having read the passage as a separate classification from either of the two comprehension systems.

The categorizing of the text passages and the items was verified by reproducibility, the degree to which the process can be replicated by multiple researchers (Krippendorff, 2004). I trained a colleague on the classification techniques by working through two samples together, one from the NAEP and one from a state test. She was asked to identify the test items as multiple choice or constructed response. Next, items were classified based on the NAEP categories. Finally, items were classified based on the level of comprehension categories. After working through the examples and checking for understanding of the process, my colleague duplicated the coding process on a sample of ten percent of the related items. The process was considered successful when a minimum of 90% agreement was reached between the two analysts. (Krippendorff, 2004).

Analysis

Descriptive statistics were examined for each of the variables. The data were checked for accuracy and normality. Each variable was examined for the states alone, as well as for the entire sample including the NAEP. Correlations were computed to examine the potential relationships between any of the given pairs of independent variables. Pearson coefficients were examined to find significant correlations.

The results of the measures of difficulty for each passage were entered into a database. Stepwise regression was selected to determine if any of the independent variables predicted the dependent variable of difference in proficiency percentages between the 2005 state reading assessments and the 2005 NAEP elementary reading test (Spicer, 2005). The independent variables were per pupil spending, length of text

passage, Spache readability result, Powers, Sumner, Kearsley readability result, inclusion of an introduction, inclusion of an illustration, percentage of constructed response test items, percentage of multiple choice test items, percentage of test items for each of the NAEP aspects of reading, percentage of test items for each of the Sadoski comprehension levels, and percentage of test items that could be answered without reading the text passage. The independent variables were entered in several groups based on text difficulty factors and types of comprehension questions. The stepwise regression results for each group were examined for the order that the variables were entered, the change in R^2 with the entry of each variable, the total R^2 , the significant change of R^2 , the significant beta weights and the total amount of variance accounted for in the model (Spicer, 2005; Grimm & Yarnold, 1998). SPSS 15 software was used for all of the descriptive statistics, correlations and regression computations.

Examining the comparisons between the state assessments and between the state assessments and the NAEP required the use of a special analysis procedure due to the small sample size within each group, the NAEP, the per pupil spending quartiles and the performance level groups determined by the differences between the proficiency percentages of fourth grade students on the 2005 state reading test and the NAEP reading test. Due to the sample sizes in these groups being between four and fifteen, traditional parametric procedures were ruled out. Bootstrapping (Efron & Gong, 1983; Rodgers, 1999) is a method of estimating the parameters of a given population based on the actual samples observed in the research. The bootstrap duplicates the population a large number of times using resampling with replacement to produce a bootstrapped mean and estimates of the confidence intervals. Group comparison is made by comparing the

confidence intervals. If the confidence intervals overlap between groups, no significant difference exists between the groups. If the confidence intervals do not overlap, the groups are found to differ significantly (Efron & Gong, 1983, Rodgers, 1999). The bootstrapping procedures were completed with Resampling Procedures software retrieved from <http://www.uvm.edu/~dhowell/StatPages/Resampling/Resampling.html>.

Chapter 4

Findings

The goal of this study was to determine if differences exist between tests used under the NCLB accountability system. I examined released and sample test passages and the related test items from the fourth grade reading tests from twenty-eight states as well as from the NAEP to determine how 2005 state elementary reading assessments compare to one another and to the NAEP.

Variables

Table 3 provides the descriptive statistics of mean, standard deviation, median, and range for the variables affecting differences between the states in the sample, the per pupil spending of each state and the difference between the state and the NAEP fourth grade reading tests for each of the twenty-eight states in the sample. The samples of per pupil spending and performance difference on the 2005 state tests did not appear to

Table 3

Descriptive statistics:	State factors
Variable	Mean (Standard Deviation) Median Range
Per Pupil Spending	\$8,216.19 (\$1,744.25) \$7,748.00 \$6,036-\$12,981
Difference on '05 tests*	40.54 (15.80) 44.00 6-71

*reported as the difference in percentage of students scoring proficient or above on the state test and on the NAEP

violate assumptions of normality. However, the data for the variable of per pupil spending is more clustered at the lower end than the higher end, resulting in a greater range above the median (\$7,749-\$12,981) than below the median (\$6,036-\$7,748). The opposite is true of the difference on the 2005 state and NAEP tests. The data are more clustered at the top end of the range, drawing the median above the mean by four points. The range at the lower end is 38 (6-44), while at the upper end of the scale the range is only 27 (44-71).

Table 4 displays the descriptive statistics for the different types of test items. The descriptive statistics for the variables regarding the types of items show that the NAEP includes a greater percentage of constructed response test items than the state tests, and the state tests include a greater percentage of both multiple choice items and items that can be answered without reading the text passage. The NAEP does not include any test items that can be answered without reading the accompanying passage.

Table 5 presents descriptive statistics for variables affecting text difficulty. Note that the passage length is longer on the NAEP than on the state tests. The readability figures are slightly higher for the NAEP than the state assessments. The Powers, Sumner, Kearsley readability formula, based on sentence length and number of syllables in a 100 word sample of text, used in this study found ranges in grade equivalent levels of 4.7-5.8 on the NAEP samples and 3.7-6.4 on the state samples. The Spache readability formula, based on sentence length and the use of unfamiliar words found ranges from 3.1-4.2 for the NAEP and 2.3-6.4 for the state test samples. The inclusion of an introduction to the text differed, with the NAEP not including any introductions and forty-four percent of the

state passages having one. Both the NAEP and state tests included illustrations with some of the passages, with the states including illustrations with more passages than the NAEP.

Table 6 presents the descriptive statistics for the NAEP comprehension categories, as well as a category, NAEP Categories Do Not Apply, that had to be added during the study for questions that did not fit any of the NAEP comprehension categories. Some items on the state tests could not be classified into any of the existing NAEP comprehension categories. Note that the NAEP included higher percentages of two of the categories, Developing Interpretation and Making Reader-Text Connections, while the state tests had higher percentages of items in the other two categories, Forming General Understanding and Examining Content and Structure.

Table 4

Descriptive statistics: Test Item Factors

Variable	NAEP	States
	Mean	Mean
	(St. Deviation)	(St. Deviation)
	Median	Median
	Range	Range
Constructed Response Items*	.59 (.10) .59 .50-.67	.14 (.17) .11 .00-1.0
Multiple Choice Items*	.42 (.10) .42 .33-.50	.86 (.17) .89 .00-1.0
Items answered w/o passage*	.00 (.00) .00 .00-.00	.05 (.11) .00 .00-.50

*reported as the ratio of item type to total items

Table 5

Descriptive statistics: Factors Affecting Passage Difficulty

Variable	NAEP	States
	Mean	Mean
	(Standard Deviation)	(Standard Deviation)
	Median	Median
	Range	Range
Passage length ^a	882.75 (219.84) 819.50 694-1198	535.61 (234.15) 530.00 65-1017
Spache readability ^b	3.7 (.50) 3.75 3.1-4.2	3.3 (.59) 3.2 2.3-6.4
PSK Readability ^c	5.30 (.58) 5.35 4.7-5.8	5.03 (.51) 4.90 3.7-6.4
Introduction to text ^d	.00 (.00) .00 0-0	.44 (.50) .00 0-1
Illustration included ^e	.50 (.58) .50 0-1	.69 (.47) 1.0 0-1

a: total number of words per passage, including title and introduction, if included

b: grade level equivalent (grade level & month) as computed with Spache readability formula

c: grade level equivalent (grade level & month) as computed with Powers, Sumner, Kearsley readability formula

d: reported as the ratio of passages that included an introduction to total number of passages in NAEP and state groups

e: reported as the ratio of passages including an illustration to total number of passages in NAEP and state groups

Table 6

Descriptive statistics: NAEP Comprehension Categories

Variable	NAEP	States
	Mean	Mean
	(St. Deviation)	(St. Deviation)
	Median	Median
	Range	Range
NAEP Categories Do Not Apply*	.00 (.00) .00 .00-.00	.12 (.16) .00 .00-.60
General Understanding Items*	.10 (.01) .11 .08-.11	.15 (.20) .10 .00-1.0
Developing Interpretation Items*	.69 (.10) .67 .60-.83	.53 (.23) .50 .00-1.0
Reader-Text Connections Items*	.13 (.10) .11 .08-.20	.05 (.08) .00 .00-.25
Examining Content & Structure Items*	.08 (.05) .11 .00-.11	.15 (.16) .12 .00-.83

*reported as ratio of item type to total number of items per passage

Table 7 presents the descriptive statistics for the NAEP and state assessments for the alternate comprehension categories for addressing the higher order thinking requirements of the assessment items. The literal understanding category is one that is not addressed in the NAEP comprehension categories, and is reflected here as none of the

NAEP items were categorized as measuring literal understanding. The NAEP also did not include any Critical Reading test items, and the state sample included very few. Neither the NAEP nor the state assessments included any Appreciation items, so this category was not included in any further analysis since there are no examples of these items in the samples.

Table 7

Descriptive statistics: Traditional Comprehension Categories

Variable	NAEP	States
	Mean	Mean
	(St. Deviation)	(St. Deviation)
	Median	Median
	Range	Range
Literat Understanding Items*	.00 (.00) .00 .00-.00	.06 (.11) .00 .00-.40
Inferential/Interpretation Items*	.49 (.11) .53 .33-.58	.63 (.25) .67 .00-1.0
Critical Reading Items*	.00 (.00) .00 .00-.00	.01 (.04) .00 .00-.20
Application Items*	.51 (.11) .47 .42-.67	.30 (.24) .24 .00-1.0
Appreciation Items*	.00 (.00) .00 .00-.00	.00 (.00) .00 .00-.00

*reported as ratio of item type to total number of items per passage

Variable Correlations

Correlations of the variables were computed to determine if relationships exist between the variables. Significant low to moderate positive correlations were found between the difference in proficiency percentages between the tests and Forming a General Understanding items as well as Test Items that can be answered Without the Passage, suggesting that states with tests with a greater difference in the proportion of students scoring proficient or above with the NAEP are more likely to have higher percentages of Forming a General Understanding items and Items that can be answered without reading the Passage on their assessments. Significant low to moderate positive correlations were found between passage length and the inclusion of an introduction, inclusion of an illustration, Developing Interpretation Test Items and Inferential/Interpretive Test Items, suggesting that test passages that are longer in length are more likely to have these characteristics and types of items associated with them. The Spache and Powers, Sumner, Kearsley readability formulas were correlated with each other, producing a moderate to high significant positive correlation, suggesting that the readability results align with each other by showing lower and higher readability rates similarly. These two readability formulas did not significantly correlate with any other variables.

For the types of test items, the Constructed Response test items produced significant low to moderate positive correlations with Forming a General Understanding items, Making Reader/Text Connections Items, and Application Items, suggesting that on tests with more of these types of test items, there are likely to be higher percentages of Constructed Response Test Items. The Multiple Choice Test Items produced significant

low to moderate positive correlations with Developing Interpretation Test Items and Inferential/Interpretation Test Items, suggesting that assessments that include higher percentages of these types of questions are likely to have a greater percentage of Multiple Choice Test Items on the test. The Test Items that Did Not Apply to the NAEP Comprehension Categories produced moderate to high significant positive correlations with Test Items that could be answered without Reading the Passage and Literal Understanding items, suggesting that state assessments that include questions that could not be classified under the NAEP comprehension categories were more likely to have literal level test items and items that could be answered without reading the passage. Remaining types of questions that produced significant positive correlations with each other were moderate to high Inferential/Interpretive with Developing Interpretation, and low to moderate correlations for Critical Reading with Examining Content and Structure and Application with Examining Content and Structure. These pairings suggest that as one of the types of questions occurs on a state assessment, it is likely that the other type of test item would be likely to be included as well.

A significant low to moderate negative correlation was found between the difference in the proficiency levels of the tests and the per pupil spending, suggesting that the higher the difference in the test scores between the state test and the NAEP, the more likely it is that less money was spent per child in the state for education. The difference in the proficiency level also produced low to moderate significant negative correlations with the passage length, inclusion of an introduction and the inclusion of an illustration, suggesting that states with a larger difference in proficiency percentages on the tests are more likely to have shorter test passages without introductions or illustrations. In

Table 8: Correlations, State Variables

Variable	A	B	C	D	E	F	G
A-\$	-						
B-Diff.	-.351*	-					
C-length	.081	-.302*	-				
D-Spach	.012	-.046	-.098	-			
E-PSK	-.075	-.129	.030	.679**	-		
F-CR	.098	-.030	-.260	-.061	-.090	-	
G-MC	-.098	.030	.260	.061	.090	-1.0	-
H-Intro	.169	-.523**	.381**	.115	.246	.215	-.215
I-Ill.	.124	-.305*	.323*	.026	.128	.066	-.066
J-w/o	-.271	.334*	-.207	-.062	-.207	-.220	.220
K-DNA	-.116	.065	-.156	.054	-.095	-.118	.118
L-GU	-.036	.345*	-.306*	-.157	-.162	.396**	-.396**
M-DI	.064	-.189	.386**	.023	.139	-.366**	.366**
N-RT	.161	-.253	.084	.006	-.107	.327*	-.327*
O-CS	.002	-.102	-.009	.137	.208	-.032	.032
P-LU	.198	-.147	-.049	.229	.188	.046	-.046
Q-Inf	.093	.006	.312*	-.149	-.120	-.346*	.346*
R-CRdg	-.019	-.072	-.058	.147	.108	.096	-.096
S-Appl	-.165	.073	-.294*	.025	.018	.320*	-.320*

Table 9: Correlations, State Variables, continued

	H	I	J	K	L	M
H-Intro	-					
I-Ill.	.258	-				
J-w/o	-.223	-.190	-			
K-DNA	-.128	-.012	.616**	-		
L-GU	-.113	-.210	-.087	-.209	-	
M-DI	.104	.267	-.052	-.197	-.581**	-
N-RT	.085	.043	-.253	-.238	-.116	-.225
O-CS	.068	-.109	-.327*	-.356**	-.135	-.398**
P-LU	.074	.171	-.031	.629**	-.182	-.092
Q-Inf	-.091	.142	-.176	-.444**	-.114	.619**
R-CRdg	.096	-.049	-.056	-.121	-.138	-.003
S-Appl	.043	-.212	.204	.190	.220	-.591**

Key for correlation tables:

A=\$	Per Pupil Spending
B= Diff.	Difference in proficiency percentages on 2005 4 th grade state reading tests and NAEP
C=length	Passage length measured in number of words
D=Spach	Spache readability formula results
E=PSK	Powers, Sumner Kearsley readability formula results
F=CR	Constructed Response test items, measured in percentage of total test items
G=MC	Multiple Choice test items, measured in percentage of total test items
H=Intro	Introduction to the text passage, measured in percentage of passages including intro.
I=Ill.	Illustration included, measured as one or more illustrations or graphics included with text
J=w/o	Test item could be answered without reading the accompanying text passage
K=DNA	NAEP categories Did Not Apply to classifying the comprehension item
L=GU	Forming a General Understanding, NAEP category for comprehension
M=DI	Developing Interpretation, NAEP category for comprehension
N=RT	Making Reader/Text Connections, NAEP category for comprehension
O=CS	Examining Content and Structure, NAEP category for comprehension
P=LU	Literal Understanding, traditional category for comprehension
Q=Inf	Inferential/Interpretive, traditional category for comprehension
R=CRdg	Critical Reading, traditional category for comprehension
S=Appl	Application, traditional category for comprehension

Table 10: Correlations, State Variables, continued

	N	O	P	Q	R	S
N-RT	-					
O-CS	.210	-				
P-LU	-.115	-.228	-			
Q-Inf	-.112	-.230	-.266	-		
R-CRdg	-.116	.352*	-.144	-.073	-	
S-Appl	.191	.280*	-.158	-.896**	-.031	-

addition, passage length also produced significant low to moderate negative correlations with Forming a General Understanding Test Items and Application Test Items, suggesting that states with longer text passages on the assessments are more likely to have fewer of these types of test items. Constructed Response Test Items and Multiple Choice Test Items are significantly negatively perfectly correlated at -1.0. The percentage of constructed response items also produced significant low to moderate negative correlations with Developing Interpretation and Inferential/Interpretive test items. The percentage of Multiple Choice Test Items on the assessments produced significant low to moderate negative correlations with Forming a General Understanding Test Items, Making Reader-Text Connections items, and Application items, suggesting that as the percentage of multiple choice test items increases, the inclusion of General Understanding, Making Reader-Text Connections, and Application test items decrease. The variable for Test Items that Do Not Apply to the NAEP category produced low to moderate significant negative correlations with Examining Content and Structure and

Inferential/Interpretive Test Items, suggesting that as the percentages of test items that do not fit the NAEP standards increase, these two other types of questions are less likely to be included on the state tests. Inferential/Interpretive test items and Application test items were significantly highly negatively correlated, suggesting that as the percentage level of one of these types of test item increases, the other decreases. The Developing Interpretation Test Items produced a significant low to moderate correlation with Examining Content and Structure and significant moderate to high negative correlations with Forming a General Understanding and Application Test Items, suggesting that states with higher percentages of Developing Interpretation Test Items on their assessments are likely to have lower percentages of these three other types of test items. The last pairs of significant correlations are low to moderate negative correlations between Examining Content and Structure and Test Items that Do Not Apply to the NAEP comprehension categories and Test Items that can be answered without reading the associated text passage, suggesting that states that have higher percentages of Examining Content and Structure Test Items are more likely to have lower percentages of these two types of test questions.

Comparisons of state tests

The first part of each research question addressed whether differences exist between state tests. These portions of the research questions were addressed by using stepwise regression to determine which of the independent variables had a significant effect on the criterion variable of the difference in the percentage of students reaching proficiency or above between the 2005 NAEP and the 2005 state fourth grade reading

tests. Three different combinations were examined to determine which, if any, variables help to predict the difference between the state and NAEP proficiency levels.

Table 11 presents the results of the first group of stepwise regression calculations. The Model 1 attempts used factors related to the difficulty of the reading task as independent variables. Models 1a and 1b included Per Pupil Spending, Passage Length, the inclusion of Illustrations and Introductions, and one of the readability formulas as the independent variables. Model 1a included the Spache Readability results and Model 1b included the Powers, Sumner, Kearsley readability results. Both versions of the model had the same results. Introduction to the Text was the first variable entered, accounting for 27% of the variance in the difference of the proficiency percentages on the 2005 state and NAEP scores. The next variable entered was Per Pupil Spending, accounting for another 7% of the difference in proficiency levels, for a total of 34% of the variance for each of the versions of the model. The excluded variables were the passage length, the inclusion of illustrations with the text, and the readability results. Model 1c added the types of test items, Constructed Response and Multiple Choice to the independent variables, with identical results as Models 1a and 1b. Model 1d included the independent variable of Test Items that could be answered without reading the Text, and produced the same results as the other models in this section.

These results suggest that the biggest predictor with these groups of variables is whether or not an introduction is included with the text. The significant results for including the Introduction to the Text and the Per Pupil Spending variables were negative, suggesting that the fewer passages that include introductions to the text and the less money that is spent per pupil in a state may be predictors of a greater difference in

Table 11: Stepwise Regression Model 1

Model Variable Entered	B (constant) (standard Error)	
	Beta (Standard Error)	R ² R ² Change
Models 1a, 1b, 1c, 1d Step 1	47.75 (2.60)	
Intro to Text	-.52** (3.86)	.27 .27
Step 2	67.43 (8.88)	
Intro to Text	-.47** (3.77)	
Per Pupil Spending	-.28* (.001)	.34 .07

*Significant at $p < .05$ ** Significant at $p < .01$

the proficiency rates of fourth grade students on the state test and the NAEP. These results also suggest that the other independent variables of Passage Length, Inclusion of Illustrations, readability figures, Constructed Response Items, Multiple Choice Items, and inclusion of Test Items that can be answered without Reading the Text do not help to predict the difference in the proportion of students scoring proficient or above between the state and NAEP fourth grade reading assessments. Model 2 included Per Pupil Spending and Inclusion of an Introduction, since these were already found to be predictive of the differences in the test scores, variables that affected test difficulty of Length of Passage, Constructed Response Items, and Multiple Choice Items, as well as the measures of comprehension used on the NAEP. These results are summarized in

Table 12: Stepwise Regression Model 2

Model Variable Entered	B (constant) (standard Error)	
	Beta (Standard Error)	R ² R ² Change
Model 2		
Step 1	47.74 (2.60)	
Intro to Text	-.52** (3.86)	.27 .27
Step 2	43.66 (2.95)	
Intro to Text	-.48** (3.70)	
General Understanding	.30* (9.11)	.36 .09
Step 3	63.22 (8.52)	
Intro to Text	-.43** (3.58)	
General Understanding	.29* (8.67)	
Per Pupil Spending	-.27* (.001)	.43 .07

*Significant at $p < .05$ ** Significant at $p < .01$

Table 12. The Length of Passage, Constructed Response Items and Multiple Choice Items were included to see if these variables might help to predict the score difference in conjunction with the NAEP comprehension categories. The first variable entered was again the Inclusion of an Introduction to the Text, accounting for 27% of the variance in the difference of test scores, with a negative relationship to the dependent variable, suggesting that the fewer the number of text passages that include introductions on a state test, the more likely the results are to have a large difference from the NAEP results. The next variable entered was Forming a General Understanding, which accounted for an additional 9% of the variance in the dependent variable, for a total of 36%. This suggests that states with a greater number of Forming a General Understanding questions are more likely to have state fourth grade reading test results with larger differences from the NAEP results. The final variable entered was Per Pupil Spending, accounting for an additional 7% of the variance, for a total of 43% of the variance in the difference in proficiency percentages for Model 2. The excluded variables were the Constructed Response Items, Multiple Choice Items, and the other three types of NAEP comprehension categories, Developing Interpretation, Making Reader/Text Connections, and Examining Content and Structure, suggesting that these independent variables do not contribute to the prediction of the difference in the test scores.

The attempts at the third model included the traditional comprehension categories in combinations with other variables. The results of the Model 3 attempts are summarized in Table 13. Model 3a included traditional comprehension categories along with the Inclusion of an Introduction to the Text and Per Pupil Spending since they had already been found to be predictive of the dependent variable, as well as the Constructed

Response Items, Multiple Choice Items and Length of Passage to see if these variables might interact with the new group of comprehension variables. Model 3b included the Spache readability results to see if including that variable with a new group of independent variables might affect its inclusion in the prediction model. Model 3c excluded Per Pupil Spending to see if taking that variable out of the group of independent variables would have any affect on the inclusion of other independent variables. The inclusion of the readability formula and the exclusion of Per Pupil Funding had no effect on other variables being included and resulted in the same results as Model 3a, except that for Model 3c there was only one included variable, so the results stop after Step 1, the inclusion of the Introduction to the Text.

Table 13: Stepwise Regression Model 3

Model Variable Entered	B (constant) (standard Error) Beta (Standard Error)	R ² R ² Change
Model 3a, 3b Model 3c (Step 1 only) Step 1	47.85 (2.67)	
Intro to Text	-.52** (3.94)	.27 .27
Step 2	70.31 (9.55)	
Intro to Text	-.44** (3.87)	
Per Pupil Spending	-.30* (.001)	.35 .08

*Significant at $p < .05$ ** Significant at $p < .01$

The results from the Model 3 attempts suggest that none of the traditional comprehension categories have any role in predicting the difference in proficiency percentages between the fourth grade state tests and the NAEP. These models found the same independent variables predicting the variation in the scores, although with slightly different results than the other models. Model 3 found that the Inclusion of the Introduction to the Text again accounted for 27% of the variance in the test scores, but that the addition of the Per Pupil Spending in this combination of variables added 8% more variance, for a total of 35%.

Comparison between states and between state and NAEP tests

The comparisons between the passages and items on the state assessments and the NAEP and between the groups of states were computed using the bootstrap technique. While bootstrapping is a less powerful measure than traditional parametric statistical methods, it identifies differences between groups by finding the confidence intervals for the bootstrapped means. If the confidence interval ranges overlap between the groups, there is no significant difference. If the ranges do not overlap, the two groups have significant differences.

Table 14 presents the comparisons of the means and the confidence intervals for the variable of passage length across quartiles and performance levels and the NAEP. The results for the variable of Passage Length show that significant differences exist between the NAEP passage length and the length of passages used in tests in states in Quartiles 2, 3 and 4. Additionally, when all of the state tests are examined together, significant differences exist between the length of passages for the state tests and the NAEP. The NAEP passage length does not differ significantly from the passage length

for states in Quartile 1, suggesting that states that spend more money per pupil also have elementary reading assessments with passages closer in length to those used on the fourth grade NAEP. No significant differences were found between the groups of states based on the quartiles of per pupil funding for the variable of passage length.

The NAEP passages also had similarities in length with states in Performance Levels 1 and 2, suggesting that states in these groups with state elementary reading scores most closely aligned with the NAEP scores used reading passages of similar length on their assessments. There were significant differences between the NAEP reading passage lengths and the length of the passages on the state tests in Performance Levels 3 and 4, suggesting that there are differences in the length of the passages as the difference in the proportion of students scoring proficient or above becomes greater between the state tests and the NAEP. No significant differences were found between the states in the Performance Level groups based on the length of the reading passages on the assessments. Table 15 summarizes the bootstrapped comparisons for one of the measures of passage difficulty, the Spache readability formula, and Table 16 summarizes the results for the Powers, Sumner, Kearl readability formula. The results for both of these variables show that each of the confidence intervals overlap, suggesting that there are no significant differences between any of the state groups based on quartiles of per pupil funding or performance levels and the NAEP based on either of the readability formulas used, the Spache or the Powers, Sumner, Kearl. The overlap of the confidence intervals within each group also suggests that there are no significant differences between the groups of states based on either of the readability formula results for the groups of states based on per pupil spending or performance levels. Additionally, when looked at as a

Table 14

Passage Length (number of words per passage) Bootstrapped Comparisons

a-difference with NAEP b-difference with other Quartiles c-difference with other Performance Groups











<u>Group</u>	<u>BS Mean</u>	<u>Standard Error</u>	<u>CI Lower</u>	<u>CI Upper</u>	<u>Comparison Graph</u>	
					300	1100
Q1	603.54	50.86	508.48	760.36		—
Q2	513.89 ^a	69.60	349.05	669.38	—	
Q3	515.50 ^a	74.57	315.74	674.34	—	
Q4	542.81 ^a	53.67	430.45	711.41	—	
All States	529.80 ^a	32.92	469.21	601.13	==	
NAEP	878.33	91.24	733.25	1095.50		===
P1	709.39	67.51	549.53	928.80	
P2	539.95	60.30	413.13	734.62	
P3	511.69 ^a	47.69	422.74	640.57	
P4	474.41 ^a	61.66	343.30	617.54	

Table 15: Spache Readability (Grade Equivalent) Bootstrapped Comparisons
a-difference with NAEP b-difference with other Quartiles c-difference with other Performance Groups

<u>Group</u>	<u>BS Mean</u>	<u>Standard Error</u>	<u>CI Lower</u>	<u>CI Upper</u>	<u>Comparison Graph</u>
					3.0 5.0
Q1	3.53	.16	3.20	4.01	
Q2	3.06	.15	2.74	3.47	
Q3	3.22	.14	2.91	3.67	
Q4	3.57	.14	3.22	3.87	
All States	3.31	.08	3.14	3.48	
NAEP	3.70	.22	3.30	4.10	
P1	3.47	.15	3.11	3.80	
P2	3.14	.16	2.79	3.63	
P3	3.37	.16	3.11	3.88	
P4	3.28	.15	2.97	3.63	

Table 16:

Powers Summer Kearl Readability (Grade Level Equivalent) Bootstrapped Comparisons

<u>Group</u>	<u>BS Mean</u>	<u>Standard Error</u>	<u>CI Lower</u>	<u>CI Upper</u>	<u>Comparison Graph</u>	
					4.0	6.0
Q1	5.18	.11	4.87	5.51		
Q2	4.81	.12	4.54	5.10		
Q3	5.03	.18	4.67	5.60		
Q4	5.23	.15	4.91	5.67		
All States	4.80	.13	4.58	5.13		
NAEP	5.29	.26	4.80	5.80		
P1	5.29	.14	4.96	6.13		
P2	4.99	.14	4.69	5.44		
P3	5.29	.12	4.98	5.53		
P4	4.90	.12	4.59	5.12		

whole, the group of all states did not differ significantly from any of the other state groups or the NAEP passages for either of these variables.

Table 17 presents the results for the bootstrapped comparisons for the variable of constructed response test items. The ratio of constructed response test items to total test items for the NAEP differs significantly with each of the other groups. The NAEP has significant differences with the ratio of constructed response items for each group of states based on per pupil funding and performance levels, as well as with the group of all states for this variable. While very slight differences could exist between two of the quartiles for per pupil funding, the difference is so small that a significant difference is most likely not suggested. Quartiles 2 and 4 have a 0.01 difference in the ratio of constructed response test items to total test items, but a difference of that margin most likely is not suggesting an important difference between the groups.

Table 18 shows the ratio of multiple choice response test items to total test items for the NAEP differs significantly with each of the other groups. The NAEP has significant differences with the ratio of multiple choice response items for each group of states based on per pupil funding and performance levels, as well as with the group of all states for this variable. Quartile 2 and 4 have a 0.02 difference in the ratio of multiple choice test items to total test items, but a difference of such a small margin most likely is not suggesting an important difference between the groups. Additionally, while very slight differences could exist between Quartile 4 for per pupil funding and the group of All States, the difference is so small (.01) that a significant difference is most likely not suggested.

Table 17: Constructed Response Test Items Bootstrapped Comparisons

a-difference with NAEP b-difference with other Quartiles c-difference with other Performance Groups

Group	BS Mean	Standard Error	CI Lower	CI Upper	Comparison Graph
Q1	.14 ^a	.03	.06	.20	
Q2	.27 ^{ab}	.06	.11	.45	
Q3	.17 ^a	.05	.08	.25	
Q4	.05 ^{ab}	.02	.01	.10	
All States	.16 ^{ab}	.02	.10	.21	
NAEP	.59	.04	.50	.67	
P1	.13 ^a	.02	.09	.20	
P2	.16 ^a	.04	.07	.25	
P3	.14 ^a	.04	.08	.24	
P4	.13 ^a	.06	.03	.26	

Table 18: Multiple Choice Test Items Bootstrapped Comparison

a-difference with NAEP b-difference with other Quartiles c-difference with other Performance Groups

Group	BS Mean	Standard Error	CI Lower	CI Upper	Comparison Graph
Q1	.87 ^a	.03	.81	.93	
Q2	.76 ^{ab}	.06	.58	.89	
Q3	.84 ^a	.04	.73	.94	
Q4	.95 ^{ab}	.02	.91	.99	
All States	.82 ^{ab}	.03	.75	.89	
NAEP	.42	.04	.33	.50	
P1	.86 ^a	.02	.80	.93	
P2	.85 ^a	.04	.76	.94	
P3	.86 ^a	.04	.76	.93	
P4	.81 ^a	.06	.63	.96	

Table 19 shows that all of the confidence intervals for the NAEP comprehension variable of Forming a General Understanding confidence intervals overlap. This suggests that there are not any significant differences between any of the groups of states, including the entire group of states together, and the NAEP assessments for this variable. There also are not significant differences between any of the quartiles or between any of the performance levels based on the variable of comprehension questions included on the assessments that are classified as Forming a General Understanding.

Table 20 presents the bootstrapped comparison data for the variable of NAEP comprehension for Developing Interpretation, which suggest a slight difference (.02) between the ratio of Developing Interpretation test items as compared to total test items between the group of all states and the NAEP tests. This small difference most likely does not represent an important difference between the Developing Interpretation items in the group of state tests and the NAEP test. A significant difference was suggested between the NAEP and Performance Level 4, suggesting that as the gap between the performance levels on the state elementary reading assessments and the NAEP increases, that there is a significant difference between the ratio of Developing Interpretation comprehension test items included on the reading assessment. The confidence intervals between the groups of states in the quartile groups and the groups of states in the performance level groups overlap, suggesting that there are no significant differences between the states in these groups based on their use of comprehension questions for developing interpretation.

Table 21 presents the results for Making Reader/Text Connections. The results show that Quartile 4 for per pupil spending is significantly different than the NAEP items

for Making Reader/Text Connections, suggesting that states that spend less per pupil for education create assessments with fewer test items that measure Making Reader/Text Connections. There is also a slight (.01) difference between the group of states as a whole and the NAEP scores.

There was no significant difference between the NAEP and Quartiles 1, 2, and 3 for per pupil spending, suggesting that states in these groups use similar ratios of this type of question. There also was no significant difference between the NAEP and the states in performance levels 1 and 2, suggesting that states in these groups with scores closest to those of the NAEP have similar ratios of Making Reader/Text Connection test items on their assessments. The states in Performance Levels 3 and 4 had significant differences from the NAEP for the variable of test items measuring Making Reader/Text Connections. This suggests that states whose proficiency levels are farther from those on the NAEP for their state use a smaller percentage of Making Reader/Text Connections on their elementary reading assessments. The groups of states did not have any significant differences between the quartiles or performance level groups based on the use of comprehension questions to develop Reader/Text connections on the assessments.

Table 22 presents the bootstrapped comparison information for the variable of Examining Content and Structure. The results show that all of the groups have overlap within their state groupings and with the NAEP for the variable of Content and Structure. This suggests that there is no significant difference between groups of states based on per pupil spending or performance levels, or between any of these groups and the NAEP in regard to test items for Content and Structure.

Table 19: General Understanding Test Items Bootstrapped Comparison

a-difference with NAEP b-difference with other Quartiles c-difference with other Performance Groups

Group	BS Mean	Standard Error	CI Lower	CI Upper	Comparison Graph
Q1	.09	.03	.03	.16	—
Q2	.30	.08	.09	.48	—
Q3	.15	.05	.04	.30	—
Q4	.16	.04	.07	.25	—
All States	.15	.03	.10	.22	==
NAEP	.10	.01	.09	.11	≡
P1	.08	.03	.03	.13
P2	.07	.03	.02	.13
P3	.11	.03	.05	.19
P4	.13	.03	.04	.20

Table 20: Developing Interpretation Test Items Bootstrapped Comparison

a-difference with NAEP b-difference with other Quartiles c-difference with other Performance Groups

Group	BS Mean	Standard Error	CI Lower	CI Upper	Comparison Graph
Q1	.63	.05	.50	.74	
Q2	.42	.07	.25	.59	
Q3	.50	.05	.38	.62	
Q4	.49	.05	.39	.63	
All States	.51 ^a	.03	.46	.59	
NAEP	.69	.04	.62	.79	
P1	.55	.05	.43	.70	
P2	.60	.05	.47	.78	
P3	.62	.04	.52	.73	
P4	.44 ^a	.06	.27	.54	

Table 23 presents the bootstrapped comparisons for the variable of test items that were able to be answered without reading the accompanying text passage. The results show overlap between Quartiles 1, 2, and 3 with the NAEP test items for the variable of test items that can be answered without reading the accompanying text passage. This suggests that the state tests in these groups use similar ratios of these items on their elementary reading assessments as the NAEP. The confidence levels for Performance Level 1 overlap with the NAEP. All of these state groups include zero in the range, which matches the fact that none of these types of questions are included on the NAEP. The only significant difference involving quartiles based on per pupil spending is for Quartile 4, suggesting that states that spend less money per pupil have a greater chance of including this type of test item on their assessment. Performance Levels 2, 3 and 4 all have differences in confidence intervals with the NAEP. These results suggest that as the performance level gap between the NAEP and the state tests increase, there is a greater likelihood of including test items which can be answered without reading the passage on the state assessments. The differences between the confidence intervals are not large, which could be a reflection of the small number of test items overall that were classified into this category. The groups of states in the quartiles and performance level groups did not have any significant differences from one another related to the content and structure comprehension questions used on the assessments.

Table 21: Making Reader/Text Connections Test Items Bootstrapped Comparisons
a-difference with NAEP b-difference with other Quartiles c-difference with other Performance Groups

Group	BS Mean	Standard Error	CI Lower	CI Upper	Comparison Graph
					.00 1.0
Q1	.09	.03	.03	.14	—
Q2	.06	.02	.02	.10	—
Q3	.06	.02	.02	.11	—
Q4	.02 ^a	.01	0.0	.04	—
All States	.05 ^a	.01	.03	.08	==
NAEP	.12	.02	.09	.18	===
P1	.10	.03	.04	.18
P2	.05	.02	.01	.10
P3	.04 ^a	.02	.01	.08
P4	.04 ^a	.02	-0.01	.07

Table 22: Content and Structure Test Items Bootstrapped Comparisons

a-difference with NAEP b-difference with other Quartiles c-difference with other Performance Groups

Group	BS Mean	Standard Error	CI Lower	CI Upper	Comparison Graph
Q1	.14	.04	.06	.28	
Q2	.24	.06	.09	.39	
Q3	.15	.03	.08	.21	
Q4	.11	.03	.05	.18	
All States	.16	.02	.11	.21	
NAEP	.08	.02	.03	.11	
P1	.20	.05	.10	.34	
P2	.11	.03	.05	.17	
P3	.12	.03	.05	.19	
P4	.21	.05	.10	.36	

Test Items that can be answered without reading the passage **Bootstrapped Comparisons**
a-difference with NAEP b-difference with other Quartiles c-difference with other Performance Groups

96

Table 24 presents the bootstrapped comparisons for the variable of Literal Understanding test items. Quartiles 1, 2, and 3 had confidence intervals with significant differences with the NAEP for literal understanding comprehension questions, although all of the differences were small (.01). These findings suggest that states spending more money per pupil are more likely to include literal understanding test items on their assessments, although the small differences suggests that the differences should not have too much importance placed on them. The confidence interval for Quartile 4 overlapped with the NAEP results, suggesting that states that spent the least amount of money per pupil included approximately the same ratio of literal understanding items as the NAEP, which does not include any questions classified as Literal Understanding. None of the quartiles based on per pupil spending differed from each other.

States grouped by performance level showed that Levels 1 and 4 both had confidence intervals that overlapped with the NAEP, suggesting that the states that had proficient scores most closely aligned with the NAEP as well as those states with the largest differences from the NAEP both had similarities with the NAEP regarding the ratio of literal understanding comprehension questions. These similarities likely included having no literal understanding questions on the test, as the overlap was with zero literal understanding questions on the NAEP. Performance Levels 2 and 3 states both had differences in confidence levels with the NAEP, suggesting that these states utilize greater numbers of literal understanding comprehension questions on their assessments. None of the performance level groups showed any significant differences from each other. In addition to the separate group differences, the group of all state assessments was significantly different from the NAEP sample for literal understanding test items.

Table 24: Literal Understanding Test Items Bootstrapped Comparison

a-difference with NAEP b-difference with other Quartiles c-difference with other Performance Groups

Group	BS Mean	Standard Error	CI Lower	CI Upper	Comparison Graph
					.00 1.0
Q1	.05 ^a	.03	.01	.12	—
Q2	.05 ^a	.03	.01	.11	—
Q3	.06 ^a	.03	.01	.13	—
Q4	.04	.02	.00	.08	—
All States	.06 ^a	.02	.03	.09	=
NAEP	.00	.00	.00	.00	
P1	.06	.04	.00	.16
P2	.10 ^a	.04	.03	.20
P3	.11 ^a	.03	.02	.19
P4	.01	.01	.00	.03	...

Table 25 presents the comparisons for Inferential/Interpretive test items. States in Quartile 1 had significantly different confidence intervals with the NAEP, suggesting that states that spend more money per pupil may use greater ratios of Inferential/Interpretive comprehension items on their assessments. Quartiles 2, 3, and 4 all had overlap with the NAEP confidence intervals, suggesting that the tests do not differ in regard to the inclusion of Inferential/Interpretive comprehension test items. There were no differences between the states in the quartile groups based on Inferential/Interpretive test questions. The group of all states together differed significantly from the NAEP, suggesting that as a whole, the state assessments used a greater ratio of Inferential/Interpretive comprehension test questions. States in Performance Levels 1, 2 and 4 all had overlap of confidence intervals with the NAEP for the variable of Inferential/Interpretive comprehension items. Performance Level 3 had significantly different confidence intervals for this variable, suggesting that states with differences in proficiency between 40% and 49% from the NAEP utilize larger ratios of Inferential/Interpretive comprehension test items on their state assessments. There were no significant differences between the groups of tests within the performance level groups.

Table 26 presents the bootstrapped comparison results for the variable of critical reading questions. The group of states as a whole group differed from the NAEP in regard to the inclusion of critical reading comprehension questions, but the difference was small (.005). All other groups overlapped with the NAEP confidence intervals, suggesting that each group uses critical reading comprehension test items in similar proportions to the NAEP. None of the groups of states of per pupil quartiles or performance levels differed from each other.

Table 25: Inferential/Interpretive Test Items Bootstrapped Comparisons

a-difference with NAEP b-difference with other Quartiles c-difference with other Performance Groups

<u>Group</u>	<u>BS Mean</u>	<u>Standard Error</u>	<u>CI Lower</u>	<u>CI Upper</u>	<u>Comparison Graph</u>
Q1	.73 ^a	.05	.62	.87	
Q2	.48	.08	.33	.68	
Q3	.72	.07	.52	.86	
Q4	.63	.05	.48	.75	
All States	.64 ^a	.04	.56	.71	
NAEP	.43	.05	.33	.53	
P1	.64	.08	.42	.84	
P2	.57	.08	.35	.73	
P3	.72 ^a	.04	.63	.81	
P4	.58	.07	.43	.74	

Table 26: Critical Reading Test Items Bootstrapped Comparisons

a-difference with NAEP b-difference with other Quartiles c-difference with other Performance Groups


<u>Group</u>	<u>BS Mean</u>	<u>Standard Error</u>	<u>CI Lower</u>	<u>CI Upper</u>	<u>Comparison Graph</u>
					
Q1	.02	.02	.00	.05	—
Q2	.01	.01	.00	.02	—
Q3	.01	.01	.00	.43	—
Q4	.01	.01	.00	.03	—
All States	.01 ^{ac}	.00	.01	.02	=
NAEP	.00	.00	.00	.07	≡
P1	.02	.02	.00	.00	
P2	.00	.00	.00	.06
P3	.02	.02	.00	.00	
P4	.01	.01	.00	.02	...

Table 27: Application Test Items Bootstrapped Comparisons

a-difference with NAEP b-difference with other Quartiles c-difference with other Performance Groups











Group	BS Mean	Standard Error	CI Lower	CI Upper	Comparison Graph
					.001.0
Q1	.18 ^{ab}	.04	.08	.24	
Q2	.45 ^b	.07	.29	.65	
Q3	.30	.06	.14	.48	
Q4	.34	.06	.20	.47	
All States	.32 ^a	.03	.24	.38	
NAEP	.51	.05	.43	.61	
P1	.31	.07	.12	.52	
P2	.36	.07	.20	.52	
P3	.19 ^a	.04	.11	.27	
P4	.39	.07	.24	.56	

Table 27 presents the bootstrapped results for Application comprehension questions on the assessments. Three groups of states had significant differences with the NAEP for the variable of application comprehension test items. The confidence intervals were lower for the states in Quartile 1 than for the NAEP, suggesting that states that spend more money per pupil also use fewer application comprehension questions on their elementary reading assessments. The other quartile groups overlapped with the NAEP, suggesting that those quartiles that spend less money per pupil are more likely to use similar ratios of application questions on their assessments. In addition to the differences with the NAEP, Quartiles 1 and 2 were significantly different from each other, suggesting that the states that spend more funding per pupil use fewer application test items than the group immediately below them in per pupil spending. The set of states as a whole group was also significantly different from the NAEP, suggesting that, in general, states are likely to use fewer application comprehension test items than are used on the NAEP.

Summary

This chapter has presented the finding for the descriptive statistics, the correlations, the stepwise regressions, and the bootstrap comparisons. These techniques were all completed in order to answer the research questions aimed at determining whether differences exist between fourth grade reading assessments used in individual states and the NAEP fourth grade reading test. Of particular interest from the correlations is information about variables that significantly correlated with the difference in the test scores. Positive correlations with the Difference in Test Scores on the 2005 state tests and NAEP were the Forming General Understanding and Items that could be answered without reading the Text Passage, suggesting that as the difference between the tests

increases, state tests are likely to have higher percentages of these two types of questions. Variables that had significant negative correlations with the Difference in Test Scores were Per Pupil Spending, suggesting that the larger the difference in proficiency percentages between the state test and the NAEP, the more likely the state is to have a lower per pupil spending average. Also negatively correlated with the Difference in Test Scores were the Inclusion of an Introduction and Illustrations, suggesting that states that have larger differences in their test scores with the NAEP are less likely to include introductions or illustrations with their text passages on their fourth grade reading assessment.

Some of the results from the correlations with the Difference in Test Scores also came through as significant variables on the regressions. Both the Inclusion of an Introduction and Per Pupil Funding were found to be negative predictors of the Difference in Test Scores, indicating that states that include fewer introductions and spend less per pupil are likely to have greater differences between their state scores and the NAEP on fourth grade reading tests. Additionally, in one regression model, Forming a General Understanding test items were found to be a positive predictor of the Difference in Test Scores, suggesting that states that use higher percentages of these questions tend to have larger differences in their state fourth grade reading scores and the NAEP scores.

Finally, the bootstrap comparisons identified variables that had significant differences between groups. The variables with the most differences between the NAEP and state tests were the Constructed Response and Multiple Choice Test Items, which both had significant differences with every group of states, including the states as a

whole. Passage Length and Literal Understanding had differences between five of the eight groups of states, and with the group of states as a whole, with the NAEP, but no differences were noted for either variable between the groups of states. While Forming a General Understanding was found to be a predictor of the difference in proficiency percentages during the stepwise regressions, there were no significant differences found between that variable between the NAEP and states or between any of the groups of states.

Chapter 5

Discussion

This research study focused on examining the NAEP and state fourth grade reading assessments to determine if differences exist between the tests. Specifically, the research questions sought to determine if there were differences in text difficulty, text length, and depth of knowledge requirements between the state fourth grade reading tests as well as between the state tests and the NAEP. To answer these questions, I collected released and/or sample text passages from 28 states as well as the NAEP, one from each Reading for Literary Experience and Reading for Information classifications, if available. The passages were analyzed for readability levels, text length, and whether or not introductions and illustrations were included with the text passage. The related test items were analyzed for whether they were multiple choice or constructed response and how they assessed comprehension by two scales, the four categories used by the NAEP, and five more traditional categories of comprehension. Stepwise regression was used to see if any of the independent variables predicted the difference in the proportion of students scoring proficient or above between the 2005 fourth grade reading state tests and the NAEP. The states were also divided into groups by on performance levels based on the difference in the proportion of students scoring proficient or above between their state and the NAEP on 2005 fourth grade reading assessments and into quartiles by states' per pupil spending. These groups were compared with each other and the NAEP to see if differences existed for any of the variables.

Differences between the NAEP and state tests were found with the Constructed Response Items and Multiple Choice Items in which the NAEP groups were found to be

significantly different from each group of states, and the group of states as a whole. The NAEP was also found to be different from a majority of state groups for the variables of Literal Understanding and Passage Length. Several variables were identified as helping to predict the score difference between the state fourth grade reading tests and the NAEP. Two models that included the Inclusion of an Introduction with text passages on the tests and Per Pupil Spending were found to account for 34% and 35% of the total variance in the test scores. A model including those same two variables as well as Forming a General Understanding test items was found to account for 43% of the variance of the difference in the test scores.

Text Difficulty

The first research question in the study asked, "Do significant differences exist between state passages and state and national passages regarding the difficulty of the passages?" Several variables were examined to determine if there were significant differences in difficulty between the state tests and between the state tests and the NAEP. These variables included the Spache and Powers, Sumner, Kearsley readability results, whether or not introductions and illustrations were included with the passages and whether the test items were constructed response or multiple choice. The findings varied on the different measures of text difficulty.

The readability formulas yielded results that correlated significantly positively with each other, but did not correlate significantly with any other variables. No significant differences were found regarding the readability results between any groups of states or the NAEP and any groups of states based on per pupil spending or performance level. Readability formulas that quantify characteristics of text such as sentence length,

vocabulary use, and sentence structure suggest that they can measure text difficulty and find differences between texts based on these factors (Klare, 1984). While the ranges in grade levels for the readability results are larger for the state samples, no significant differences were found between the states or between the state tests and the NAEP based on readability, suggesting that the NAEP and state fourth grade reading tests are using text passages within similar limits to each other.

Examination of the current sample did not provide a method for differentiating between texts based on the readability formulas chosen. While it may be tempting to limit the readability levels of texts used for assessing reading proficiency when developing tests, this study suggests that it is not a worthwhile use of resources to assess and limit readability levels above or below the grade level of the test. Part of the expectation for reading proficiently is to be able to read appropriately difficult texts. The difficulty of the text is often defined by a readability formula, but with ranges spanning four grade levels in this study, no differences were found between state groups and the NAEP, suggesting that it is not a meaningful way to separate the text passages. Skilled readers decode text automatically and read for meaning (Pressley, 2006; Samuels, 2004), but the results of the current study suggest that the vocabulary and sentence length in the passages are not factors in differentiating the tests from each other or in predicting the proportions of proficient readers. These readability formulas do not take into account other important factors related to the choice of text passages on the assessment, such as genre. The readability formulas chosen for this study were not capable of determining differences related to the format of the text features that may change between Reading for Literary Experience and Reading for Information text passages. While traditional readability

formulas measure only the surface features of the text by counting words and syllables, it may be that the differences that matter in defining text difficulty for students have more to do with text format than syllable or word count. Readers' ability to navigate through headings, graphics, and captions while reading text to gain information or by reading narrative text for understanding may be more important than simply measuring the surface features represented in readability formulas.

The other measures of difficulty, the inclusion of an introduction to the text and the inclusion of one or more illustrations, were significantly negatively correlated to the difference on the test scores, but only the inclusion of an introduction was found to be a predictor of the difference in the test scores with the stepwise regression. The inclusion of an introduction was the first variable entered in each of the three models. The correlation and regressions results were significant and negative, suggesting that larger differences in proficiency rates are more likely when these two variables, inclusion of introductions and illustrations, are used in smaller numbers on state tests. The relevance of the inclusion of an introduction before a text passage and the inclusion of illustrations along with the texts suggests that the readers' interactions with cues in the text are important in the construction of meaning (Rosenblatt, 1994). While only the introduction to the text was found to be a predictor of the difference in proficiency percentage on the test, the inclusion of illustrations may also be an important tool for readers for activating prior knowledge and making connections between their experience and the text. The fact that Introductions to the text were not used on the NAEP test samples provides for interesting comparisons. The use of both the introductions and the illustrations allows the readers to have tools at their disposal to assist with their comprehension while working

independently on the test. The inclusion of these variables has suggested higher differences in scores from the NAEP, suggesting that including introductions and illustrations with text passages could assist students with meaning construction as evidenced by higher percentages of proficient readers on assessments when these variables were included. If the inclusion of introductions and illustrations assist readers with their understanding while working on the assessment, perhaps the developers of the NAEP reading assessment would consider adding introductions to the text passages in an effort to increase the readers' opportunities for preparing to read text passages on the assessment (Pressley, 2006) and create connections between their experiences and the text (Rosenblatt, 1994).

Passage Length

The second research question asked, "Do significant differences exist between state passages and state and national passages regarding passage length?" Passage length was not found to be significantly different between groups of states based on per pupil spending or performance levels, but differences were found between the NAEP and groups of states. The NAEP was found to have similarities in passage length between the first two performance levels and quartile 1, suggesting that the states that have the scores closest to the NAEP and those that spend the most per pupil have text passages of similar length to those on the NAEP. The NAEP had the longest passages of tests in this study. Similarities between the NAEP and Performance Level 1 and 2 suggests that states that use text passages of similar length on their assessments are more likely to result in scores aligned with the NAEP. All of the other groups of states had significant differences from the passage length on the NAEP, including the group of all states. Passage length was not

found to be a predictor of the difference in the proportion of students scoring proficient or above in the stepwise regression computations, although it was significantly negatively correlated with the difference in the proportion of students scoring proficient or above, suggesting that states with a larger difference in the proportion of students scoring proficient or above from the NAEP most likely have passages that are shorter in length than those on the NAEP.

The difficulty of reading relating to the length of the text passages on the assessments may be related to other factors affecting the readers' ability to read the texts. While decoding skills are not assessed in isolation on fourth grade reading tests, students' ability to decode and read fluently come into play when reading text passages. The criterion referenced state tests are not generally timed tests, although the NAEP reading tests are timed; however, timed or untimed, the reader must read fluently enough to be able to concentrate on the comprehension rather than interrupting the comprehension to decode unknown words (Pressley, 2006; Samuels, 2004, Alexander, 2005-2006). Passage length may hold the key to keeping the readers' interest throughout the reading, encouraging the students' motivation (Allington & Cunningham, 2006) by realizing that they can handle the length of the passage. The selected passages must be chosen carefully to be on a length that seems readable to the fourth grade students, but also to include vocabulary which makes reading with automaticity (Samuels, 2004) possible, giving the reader the opportunity to construct meaning while reading (Rosenblatt, 1994; Harrison, 2004; Pressley, 2006).

Passage length was significantly negatively correlated with Forming General Understanding Items and Application Items, suggesting that the higher percentage of

these types of questions that a state uses on their test, the shorter the passages are likely to be. Passage length was significantly positively correlated with Developing Interpretation Items, Inferential/Interpretive Items and Inclusion of an Introduction and Illustrations with the text passage, suggesting that states that use higher percentages of interpretation items and include introductions and illustrations with their text passages are likely to use longer text passages on their assessments. These results suggest that while shorter test passages are likely to be associated with higher percentages of proficient readers, these proficiency levels have the greatest differences from the NAEP scores, and may also be associated with lower level cognitive questions such as forming general understanding items. Tests with longer passages were associated with higher percentages of comprehension questions requiring inference and interpretation and the inclusion of introductions and illustrations, which may assist students with making connections and constructing meaning (Rosenblatt, 1994), which could result in higher proficiency scores. Test developers for the NAEP and the state reading tests may want to consider including longer passages along with introductions, illustrations, and test items that require higher order thinking skills in order to adequately assess students' reading proficiency. Additionally, test developers need to consider the level of cognition involved in the decision to include a greater number of shorter text passages on an assessment or fewer text passages of longer length. Using shorter passages may result in having a greater total number of passages on the test, but also adds to the likelihood that only lower level comprehension questions can and will be asked since the passage may not be lengthy enough to be able to examine a topic or plot of a story in much depth. On the other hand, if test developers use fewer passages of longer length, there is a greater likelihood of

comprehension questions aimed at higher order thinking and more of a possibility that the author of the longer text could have gone into more detail and depth.

Comprehension Levels

The final research question asked, “Do significant differences exist between state assessments and state and national assessments concerning higher order thinking requirements of items compared in terms of depth of knowledge/higher order thinking requirements?” The higher order thinking skills required to answer the test questions were measured through classification of the questions by type of item and the comprehension level the item addressed. This question addresses the goal of skilled reading, being able to construct meaning when reading (Pressley, 2006; Alexander, 2005-2006). Test developers need to carefully address this portion of the process since it marks the ultimate goal of reading, and as such test items need to adequately be able to measure the identified strategies.

The question types, constructed response and multiple choice, offered the most significant differences of any variables between states and states and the NAEP. The NAEP samples were significantly different from each state group based on per pupil spending, performance level, and the group of states as a whole based on each type of question. Additionally, differences were found between quartile 2 and 4 for both types of test items. Clearly, the proportion of test items that require students to write their answers as opposed to choosing from multiple choice answers is a difference between these tests; however, neither item type was identified as a predictor of the difference in the proportion of students scoring proficient or above between the NAEP and the state fourth grade tests. The use of constructed response versus multiple choice questions on a test

directly relates to the level of cognition required to complete the task. While multiple choice test items can be designed to require some higher order skills, the constructed response items require that a student organize their thoughts and be personally involved with construction of the answer. Constructed response items take away the possibility of guessing between answer choices and require that a student truly demonstrates what they know. Question types may also be integral in determining the information remembered by readers. The results of a study of fifth grade students suggested that students remembered information from expository text based on the type of questions that were asked following reading, implicit or explicit (Wixson, 1984). However, all of the questions in the study were constructed response, so while it is possible to include a range of cognitive requirements in constructed response items, perhaps the key to the students' response and recall lies not just in the cognitive level of the questions that are asked, but in the requirement of a student to become personally involved with writing the response to a constructed response item.

While the test items had correlations between variables, only Forming General Understanding was included in any of the three models as helping to predict the difference in the proportion of students scoring proficient or above between the state tests and the NAEP. There were a few differences between the state groups and the state and NAEP assessments for these test items. Differences between groups of states were evident only in the new variable of items that could be answered without reading the text passage, which had the most differences in performance levels. Performance level 1, which is the groups of states with scores closest to those on the NAEP, was different from the other groups of states, as well as different from the groups of states as a whole,

for items that could be answered without reading the text. The NAEP and those states with the most similarity of proportions of students scoring at proficient or above did not include test items with passages that could be answered without the related text passage.

Additionally, differences between the groups of states were found in the area of application test items for states grouped by Per Pupil Spending. Quartile 1 and 2 were found to be different from each other in terms of the inclusion of application test items, but were not different from the other groups of states. Quartile 1 was also significantly different from the NAEP test. These results show that the states spending the most money per pupil are more likely to include fewer application test questions on their fourth grade reading assessments. Application test items are more likely to be constructed response items as readers are asked to explain connections between their own knowledge and the text, so these findings may relate to the differences found between state groups and the state and NAEP groups regarding constructed response and multiple choice test items.

The most notable differences between the NAEP and groups of states for comprehension levels was for the categories of Items that could be answered without reading the Text Passage and Literal Understanding, both of which had four or five groups of states that were significantly different from the NAEP. The NAEP test does not include any of either of those types of questions. Most significantly, the NAEP tests were significantly different on each comprehension variable except Forming a General Understanding from the group of states as a whole. These results come together to suggest that states with results most like the NAEP in proportions of students scoring proficient or above use fewer lower level comprehension questions. Literal level

comprehension questions are the lowest cognition questions possible on the tests, and of the NAEP categories, Forming a General Understanding is at the lower end of the continuum for cognition. It is more difficult than the traditional category of literal understanding questions in that it may require a multiple step thought process in order to come to the answer. One of the steps involved may include literal level understanding, but it would be used as a step in the process to obtain the answer (National Assessment Governing Board (NAGB), 2005). The results of the analysis suggest tests that require higher levels of cognition to answer the comprehension test items end up with lower percentages of students reaching the designation of a proficient reader. This may affect test development as the NAGB, state departments of education, and assessment companies wrestle with the decisions about whether to adequately measure higher levels of comprehension and where to set the cut scores to represent the designations between satisfactory and unsatisfactory performance on the questions. To truly measure skilled reading, a test should assess a range of comprehension strategies, from literal understanding to application and interpretation of what was read and represent interaction with the text before, during and after the actual process of reading the text (Pressley, 2000; Harrison, 2004) truly represent whether a student has constructed meaning from their interactions with the text (Rosenblatt, 1994). Creating tests that rely too heavily on lower level test items does not reflect the teaching profession having high standards for itself and its consumers and does not do justice to truly measuring accountability for educators in determining if we are meeting our goal of developing skilled readers.

Implications for Policy

Information from this study has the potential to impact elementary reading assessment used across states or on the NAEP. The state fourth grade reading scores have been shown to be increasing at greater rates than the NAEP scores (NCES, 2007; Fuller & Wright, 2007). Identifying significant differences between the tests may prove useful in knowing why the test scores may or may not align, and give insight into how to construct tests that would align more closely with one another. Instead of debating which test really measures student progress, it would be possible to know what is being measured on each test, or be able to construct tests that would be in alignment to be able to measure true student learning across assessments (Amrein & Berliner, 2002). The addition of introductions to the text to prepare students prior to reading a text selection on a test seems that it would be a revision that would be possible to implement, knowing that the use of such introductions is a characteristic that creates tests with closer outcomes to the NAEP scores. An interesting point from that finding is that no introductions to the text were included with the NAEP passages.

An issue arising from the sampling in the study has implications for state departments of education across the country. During the data collection, state departments of education were contacted to verify the alignment of released and sample test passages and items to the actual fourth grade reading test. While the recent requirement date in 2006 for implementation of fourth grade reading tests may explain some of the lack of released test passages and items, it would be beneficial for school personnel to have greater access to released and/or sample tests that align with the actual fourth grade tests. Access to these released and sample passages and items may improve

the teachers' ability to prepare students adequately for the types of passages and items they will encounter on the state tests.

Further, while legislatures across the country debate whether money matters in education, this study found that per pupil spending was a predictor of differences of state reading tests with the NAEP. The less money per pupil that was spent by a state, the more likely that state was to have a larger difference in the proportion of students scoring at proficiency levels between their state test and the NAEP fourth grade reading assessment. That may be useful information in as the debate continues.

NCLB is due to be reauthorized during the current Congressional session. The first major decision that needs to be faced by Congress, or in individual state legislatures, is whether the goal of these tests used in the accountability systems should match those of the NAEP. While the NAEP continues to be considered the nation's report card, many policy makers and citizens will continue to draw comparisons whether the relationship is required by law or not. If the decision is that the tests do not need to be aligned, then policy makers and education researchers need to look for other ways to measure progress across the nation; however, if the goal is for the state tests to demonstrate growth in a similar manner to the NAEP, continuing to examine the similarities and differences in the tests will be critical to future test development, alignment and interpretation. It will be important to follow the process to see if more emphasis is placed on NAEP scores legislatively, or if it will continue to be the measuring stick for the nation without becoming the official assessment. For the moment, school accountability is built around proficiency levels on the state assessments (PL 107-110), but with each year of reported scores comes more discussion and debate about the accuracy of the scores by comparing

them to the NAEP scores (Fuller & Wright, 2007; NCES, 2007). If state departments of education truly want to prove that their students are learning, they will look for ways to align their test with the NAEP so that there will be multiple measures of student proficiency. The factors brought out in this study may be useful in finding variables that will help to align the assessments.

Limitations and Future Research

It is important to consider limitations to the current study, and how these limitations may shape the possibility of future research. One limitation is that the study could not include assessments from all of the states. Now that the deadline has passed for including fourth grade reading assessments in state accountability systems under NCLB, the study could be completed examining all of the states, however, it is still possible that due to the nature of specific state's processes that released and/or sample test items may still not be available in all cases. Another limitation of the study was that only small numbers of released and sample test passages and related items were available for many states. The small numbers of available passages and items limited the data available for analysis. When the fourth grade tests have been in use for longer periods of time it is possible that there would be greater quantities of released passages and items available, which would increase the power and robustness of the statistical analyses, and also perhaps make it possible to consider the types of text, Reading for Literary Experience and Reading for Information, separately. Type of text is an important characteristic that determines how a student approaches and completes the reading process (Duke, 2000, Pappas & Pettegrew, 1998). Further study to determine how text type affects assessment would be beneficial for future test development. As more released text passages become

available it may be possible to be able to collect and analyze the passages and items separately for the two genre classifications. This would be beneficial as it is possible that some of the factors that were not found to be indicative of differences in this study, or predictive of differences between state assessments and the NAEP, may come into play when analyzing only the Reading for Literary Experience or Reading for Information passages and related test items.

This research study endeavored to identify if differences existed between the fourth grade reading assessments. Using the knowledge of what may be different could be used as a basis to find out how these differences affect testing and how these differences measure up to the assessment frameworks for the NAEP and each state. This research study did not examine the test blueprints or test and item specifications to determine if each test matches the framework for the test in meeting its standards and difficulty levels. This research study examined the tests, but did not compare the standards being measured by each test in an effort to see if the content being assessed was similar. While the NAEP has been seen as the nation's report card, it is entirely possible that a major difference between the NAEP and the state tests is the content on which the tests are based.

This study found no significant differences in readability based upon two formulas based on word difficulty, as measured by familiarity or number of syllables, and length of sentences. It may be worthwhile to try other readability formulas to see if different parameters net different results. While the two formulas used in this research study were specifically chosen as appropriate for fourth grade texts, it may be worthwhile to look at text passages using other formulas to see whether different characteristics

would yield different results. It is also quite possible that an entirely different system of readability measurement would be more beneficial to educators and test developers. The inference load (Kemper, 1983) attempted to measure the difficulty of text in relation to readers' background knowledge. Perhaps this type of system would be beneficial to test passages in determining difficulty based on experience with the content and text format, rather than computing readability based on sentence length and vocabulary factors.

While the difference in proficiency percentages on the tests is a widely studied topic, the current study did not endeavor to examine the cut scores associated with each state test. Some of the current studies are working to equate the state scores with the NAEP (NCES, 2007), but others, including this study, compare the proportion of the proficiency rates without being able to standardize them past the designation of proficient/not proficient (Fuller & Wright, 2007). It would be worthwhile and interesting for future research to examine the cut scores for the proficiency designations, as well as the process determining how they are set in different environments.

Another limitation of this study is that Per Pupil spending was used with actual figures from across the state from the census bureau. Since Per Pupil Spending was indicated to be predictive of the difference in the percentages of proficient readers, it may be worth examining in more depth. There are several ways that this could be accomplished that would be beneficial to our understanding of the effect of per pupil spending on the assessments. First, it may be useful to equate the funding based on cost of living adjustments across the states (Carey, 2003) in order to truly make comparisons between the states. Further, it may be beneficial to examine not only the total per pupil spending, but to be able to identify the portions of that amount spent in each state on

curriculum, assessment, and professional development. The overall effect of per pupil spending may not provide a complete picture of the relationships, but being able to identify relationships between states that spend more or less in specific areas that may be directly tied to student achievement on the reading assessments may provide more insight.

Further research could examine these issues in an effort to clarify some of the differences between tests, which could affect the student proficiency outcomes. The more that is known about the test development process, cut scores, and difficulty levels of the text passages and test items, the more able we will be to develop tests to meet the demands of the task at hand and be confident that the tests will fit the purposes.

Summary

This research has focused on differences in difficulty on the NAEP and state fourth grade reading assessments. The requirements included in federal legislation (PL 110-107) mandate that all students demonstrate proficiency on state reading assessments, which are often compared to the NAEP as a snapshot of reading proficiency across the country. The research did uncover relationships between some of the variables in the study, notably significant differences between the NAEP and state test regarding the use of multiple choice and constructed response test items, as well as relationships between passage length, the difference in proficiency levels on the tests, and per pupil spending, as well as the inclusion of introductions to the text and illustrations with the text passages. Further, the study identified three models that predict the difference in proficiency levels on the NAEP and state tests, with 34% of the variance of scores attributed to the inclusion of introductions to the text passages and per pupil spending,

36% of the difference in proficiency levels attributed to the inclusion of introductions to the text and the use of forming general understanding test items, and 43% of the variance in the difference in proficiency levels attributed to all three of these variables (introduction to the text passage, forming general understanding test items, and per pupil spending) combined.

References

- A-PLUS Act (Returning NCLB Accountability to the Voters), HR 1539, 2007 Congress.
- Afflerbach, P. (2005). High stakes testing and reading assessment: National Reading Conference Policy brief. *Journal of Literacy Research*, 37(2), 151-162.
- Alexander, P.A. (2005-2006). The path to competence: A lifespan developmental perspective on reading. *Journal of Literacy Research*, 37(4), 413-436.
- Allington, R.L. & Cunningham, P.M. (2006). *Schools that work: Where all children read and write* (3rd ed.). Boston: Allyn & Bacon.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Amrein, A.L. & Berliner, D.C. (2002). High-stakes testing, uncertainty, and student learning *Education Policy Analysis Archives*, 10(18). Retrieved July 31, 2006, from <http://epaa.asu.edu/epaa/v10n18/>
- Amrein-Beardsley, A. A. & Berliner, D. C. (2003). Re-analysis of NAEP math and reading scores in states with and without high-stakes tests: Response to Rosenshine. *Education Policy Analysis Archives*, 11(25). Retrieved February 11, 2007, from <http://epaa.asu.edu/epaa/v11n25/>
- Becoming a nation of readers: The report of the commission on reading*. (1985). Prepared by Richard C. Anderson and others. Washington, D.C.: The National Institute of Education, U.S. Department of Education.

- Burger, P. & Luckman, T.. *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Garden City, NY: Doubleday, 1966.
- Berth, P. (2006). Score wars: Comparing the national assessment of educational progress with state assessments. The Center for Public Education. Retrieved October 20, 2006, from www.centerforpubliceducation.org/site/pp.aspx?c=kjJXJ5MPIsE&b=1577019
- Carey, K. (2003). *The funding gap: Low income and minority students still receive fewer dollars in many states*. Washington DC: The Education Trust.
- Carnoy, M. & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, 24(4), 305-331.
- Chall, J.S., Bissex, G.L., Conard, S.S., & Harris-Sharples, S.H. (1996). *Qualitative assessment of text difficulty: A practical guide for teachers and writers*. Cambridge, MA: Brookline Books.
- Cizek, G.J., Trent, E.R., Cranell, J., Hirsh, T., & Keene, J. (2000). Research to inform policy: An investigation of pupil proficiency testing requirements and state education reform initiatives. A presentation at the Annual Meeting of the American Educational Research Association, April 24-28, 2000, New Orleans, LA.
- Dale, E. & Chall, J.S. (1948). A formula for predicting readability. *Educational Research Bulletin*, 27, 11-20, 37-54.
- Duke, N.K. (2000). 3.6 minutes per day: The scarcity of informational texts in first grade. *Reading Research Quarterly*, 35(2). 202-224.

- Education Commission of the States. Rewards and sanctions for school districts and schools. Retrieved November 9, 2006, from <http://www.ecs.org/clearinghouse/18/24/1824.htm>
- Efron, B. & Gong, G. (1983). A leisurely look at the bootstrap, jackknife, and crossvalidation. *American Statistician*, 37(1), 36-48.
- Flesch, R. (1955). *Why Johnny can't read and what you can do about it*. New York: Harper & Row.
- Fountas, I.C. & Pinnell, G.S. (1996). *Guided reading: Good first teaching for all children*. Portsmouth, NH: Heinemann.
- Fountas, I.C. & Pinnell, G.S. (1999). *Matching books to readers: Using leveled texts in guided reading, K-3*. Portsmouth, NH: Heinemann.
- Fuller, B. & Wright, J. (2007). Diminishing returns? Gauging the achievement effects of centralized school accountability. Presentation at the American Educational Research Association, Chicago, IL, April 11, 2007.
- Graves, D.H. (2002). *Testing is not teaching: What should count in education*. Portsmouth, NH: Heinemann.
- Greene, J.P., Winters, M.A., & Forster, G. (2003). *Testing high stakes tests: Can we believe the results of accountability tests? Civic report*. NY: Manhattan Institute. Retrieved March 28, 2007, from http://www.manhattan-institute.org/html/cr_33.htm
- Grimm, L.G. & Yarnold, P.R. (1998). *Reading and understanding multivariate statistics*. Washington, DC: American Psychological Association.

- Guthrie, J.T. (2002). Preparing students for high-stakes test taking in reading. In A.E. Farstrup & S.J. Samuels (Eds.) *What research has to say about reading instruction* (3rd ed.) (pp. 370-391) Newark, DE: International Reading Association.
- Guthrie, J.T. & Kirsch, I.S. (1987). Distinctions between reading comprehension and locating information in text. *Journal of Educational Psychology*, 79(3), 220-227.
- Guthrie, J.T. & Mosenthal, P. (1986). Literacy as multidimensional: Location information and reading comprehension. *Educational Psychologist*, 22(3 & 4), 279-297.
- Hacsi, T.A. (2002). *Children as pawns: The politics of educational reform*. Cambridge, MA: Harvard University Press.
- Hanushek, E.A. (1989). The impact of differential expenditures on school performance. *Educational Researcher*, 18(4) 45-51, 62.
- Hanushek, E.A. (1994). An exchange: Part II: Money might matter somewhere: A response to Hedges, Laine, and Greenwald. *Educational Researcher*, 23(4), 5-8.
- Harris, T.L., & Hodges, R.E. (Eds.) (1995). *The Literary Dictionary: The Vocabulary of Reading and Writing*. Newark, DE: International Reading Association.
- Harrison, C. (2004). *Understanding reading development*. London: Sage Publications.
- Hedges, L.V., Laine, R.D., & Greenwald, R. (1994). Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher*, 23(3), 5-14.
- Hewitt, M.A. & Homan, S.P. (2004). Readability level of standardized test items and student performance: The forgotten validity variable. *Reading Research and Instruction*, 43(2), 1-16.

- Heubert, J.P. & Hauser, R.M. (1999). *High stakes testing for tracking, promotion, and graduation*. Washington, D.C.: National Academy Press.
- Hiebert, E.H. (2002). Standards, assessments, and text difficulty. In A.E. Farstrup & S.J. Samuels (Eds.), *What research has to say about reading instruction (3rd Ed.)*. Newark, DE: International Reading Association.
- Hoffman, J.V., McCarthy, S.J, Abbot, J., Christian, D., Corman, L., Dressman, M. et al. (1994). So what's new in the "new" basals? A focus on first grade. *Journal of Reading Behavior*, 26(1), 47-74.
- Homan, S., Hewitt, M., & Linder, J. (1994). The development and validation of a formula for measuring single-sentence test item readability. *Journal of Educational Measurement*, 31(4) 349-358.
- International Reading Association. (1999). *High stakes assessments in reading: A position statement of the International Reading Association*. Newark: DE.
Retrieved August 1, 2006, from www.reading.org
- Johnston, P.H. (1992). *Constructive evaluation of literate activity*. New York: Longman.
- Kemper, S. (1983). Measuring the inference load of a text. *Journal of Educational Psychology*, 75(3) 391-401.
- Klare, G.R. (1984). Readability. In P.D. Pearson, R. Barr, M.L. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 681-744). New York: Longman.
- Koretz, D. (2005). Alignment, high stakes, and the inflation of test scores. *Yearbook of the National Society for the Study of Education* 104 (2), 99-118.
- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology (2nd ed.)*. Thousand Oaks, CA: Sage.

- Lee, J. (2006). *Tracking achievement gaps and assessing the impact of NCLB on the gaps: An in-depth look into national and state reading and math outcome trends*. Cambridge, MA: The Civil Rights Project at Harvard University.
- Liu, D. (2006). How the federal government makes rich states richer. In The Education Trust *Funding Gaps 2006*. Washington DC: The Education Trust.
- Marchant, G. J., Paulson, S. E., & Shunk, A. (2006). Relationships between high-stakes testing policies and student achievement after controlling for demographic factors in aggregated data. *Education Policy Analysis Archives*, 14(30). Retrieved December 16, 2006, from <http://epaa.asu.edu/epaa/v14n30/>
- Mathison, S., & Freeman, M. (2003). Constraining elementary teachers' work: Dilemmas and paradoxes created by state mandated testing. *Education Policy Analysis Archives*, 11(34). Retrieved March 28, 2007, from <http://epaa.asu.edu/epaa/v11n34/>
- McQuillan, J. (1998). *The literacy crisis: False claims, real solutions*. Portsmouth, NH: Heinemann.
- MetaMetrics. (2006). Fact sheet: The Lexile framework for reading. Durham, NC: MetaMetrics. Retrieved June 12, 2007, from www.metametricsinc.com/resources/Lexile/LexileFactSheet.pdf
- National Assessment Governing Board. (2004). *Reading framework for the 2005 national assessment of educational progress*. U.S. Department of Education: U.S. Government Printing Office: Washington, D.C.
- National Center for Education Statistics (NCES) (2007). *Mapping 2005 state proficiency standards onto the NAEP scales*. Washington, DC: Institute for Education

- Statistics, US Department of Education. Retrieved June 11, 2007, from
<http://nces.ed.gov/nationsreportcard/pdf/studies/2007482.pdf>
- Oakland, T. & Lane, H.B. (2004). Language, reading, and readability formulas: Implications for developing and adapting tests. *International Journal of Testing*, 4(3), 239-252.
- Pappas, C.C. (2006). The information book genre: Its role in integrated science literacy research and practice. *Reading Research Quarterly* 41(2), 226-250.
- Pappas, C.C. & Pettegrew, B.S. (1998). The role of genre in the psycholinguistic guessing game of reading. *Language Arts* 75(1), 36-44.
- Popham, W.J. (2004). *Americas' "failing" schools: How parents and teachers can cope with no child left behind*. New York: Routledge Falmer.
- Popham, W.J. (2006). Content standards: The unindicted co-conspirator. *Educational Leadership* 64(1), 87-88.
- Pressley, M. (2006). *Reading instruction that works: The case for balanced teaching (3rd ed.)*. New York: Guilford Press.
- Public Law 89-10. 89th Congress. Elementary and Secondary Education Act.
- Public Law 107-110. 107th Congress. January 8, 2002. *No Child Left Behind Act of 2001*. Retrieved October 28, 2006, from
<http://www.ed.gov/policy/elsec/leg/esea02/index.html>
- Raphael, T.E, & Pearson, P.D. (1985). Increasing students' awareness of sources of information for answering questions. *American Educational Research Journal*, 22(2) 217-235.

- Raphael, T.E., Winograd, P., & Pearson, P.D. (1980). Strategies children use in answering questions. In Michael L. Kamil & Alden J. Moe (Eds.) *Perspectives on reading research and instruction: Twenty-ninth yearbook of the national reading conference*. 29, 56-68.
- Rasinski, T.V. (2003). *The fluent reader: Oral reading strategies for building word recognition, fluency, and comprehension*. New York: Scholastic.
- Records of the Governor and Company of the Massachusetts Bay in New England. *The Old Deluder Act* (1647). II: 203, p. 1; Retrieved March 25, 2007, from <http://personal.pitnet.net/primarysources/deluder.html>
- Rodgers, J. (1999). The bootstrap, the jackknife, and the randomization Test: A sampling taxonomy. *Multivariate Behavioral Research*, 34, 4. 441-456.
- Rodriguez, M.C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, 40(2), 163-184.
- Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context*. New York: Oxford Press.
- Rose, L.C. & Gallup, A. M. (2006). The 38th Annual Phi Delta Kappa/Gallup Poll of the Public's attitudes toward the public schools. *Phi Delta Kappan*, 88 (1), 41-56. Retrieved March 24, 2007, from <http://www.pdkintl.org/kappan/kpollpdf.htm>
- Rosenshine, B. (2003, August 4). High-stakes testing: Another analysis. *Education Policy Analysis Archives*, 11(24). Retrieved February 11, 2007, from <http://epaa.asu.edu/epaa/v11n24/>

- Sadoski, M. (2004). *Conceptual foundations of teaching reading*. New York: Guilford Press.
- Samuels, S.J. (2004). Toward a theory of automatic information processing in reading, revisited. In R.B. Ruddell & N.J. Unrau (Eds.) *Theoretical models and processes of reading: Fifth edition*. Newark, DE: International Reading Association.
- Snow, C.E., Burns, S. & Griffin, P. (Eds.), (1998). *Preventing reading difficulties in young children*. Washington, DC: National Research Council.
- Spicer, J. (2005). *Making sense of multivariate data analysis*. Thousand Oaks, CA: Sage.
- Sternberg, R.J. (1991). Are we reading too much into reading comprehension tests? *Journal of Reading*, 34(7) 540-545.
- Tierney, R.J. (2000). Literacy assessment reform: Shifting beliefs, principled possibilities, and emerging practices. In R.D. Robinson, M.C. McKenna, & J.M. Wedman (Eds.) *Issues and trends in literacy education (2nd ed.)*. Boston: Allyn and Bacon.
- U.S. Chamber of Commerce. (2007). *Leaders and laggards: A state-by-state report card on educational effectiveness*. Retrieved March 13, 2007, from <http://www.uschamber.com/icw/reportcard/default>
- U.S. Department of Education. (2007). *Building on results: A blueprint for strengthening the No Child Left Behind Act*. Washington DC: U.S. Department of Education. Retrieved June 11, 2007, from www.ed.gov/policy/elsec/leg/nclb/buildingonresults.html
- Wixson, K.K. (1984). Levels of importance of postquestions and children's learning from texts. *American Educational Research Journal*, 21(2), 419-433.

Zakaluk, B.L. & Samuels, S.J. (1988). Toward a new approach to predicting text comprehensibility in B.L. Zakaluk & S.J. Samuels (Eds.) *Readability: Its past, present, and future*. Newark, DE: International Reading Association.

APPENDIX A

APPENDIX A

Table A1: Complete State List

Selected	Quartile	State	Per Pupil \$	Perf. Diff.
	4	Alabama	\$6,553	61
	1	Alaska	\$10,114	45
X	4	Arizona	\$6,036	37
X	4	Arkansas	\$6,740	65
X	2	California	\$7,748	30
	3	Colorado	\$7,412	29
X	1	Connecticut	\$10,788	6
	1	Delaware*	\$10,228	51
X	4	Florida	\$6,784	28
X	3	Georgia	\$7,733	30
	2	Hawaii*	\$8,544	18
	4	Idaho	\$6,028	13
	2	Illinois*	\$8,656	49
	2	Indiana*	\$8,280	
	3	Iowa*	\$7,631	50
	3	Kansas*	\$7,518	50
X	4	Kentucky	\$6,888	43
X	3	Louisiana	\$7,209	30
X	1	Maine	\$9,534	34
X	2	Maryland	\$9,212	33
X	1	Massachusetts	\$10,693	55
X	2	Michigan	\$9,072	43
	2	Minnesota*	\$8,359	45
X	4	Mississippi	\$6,237	48
	3	Missouri*	\$7,331	55
X	2	Montana	\$7,763	39
	2	Nebraska*	\$8,032	26
	4	Nevada*	\$6,339	

Table A1: Complete State List, continued

Selected	Quartile	State	Per Pupil \$	Perf. Diff.
X	2	New Hampshire	\$8,860	61
X	1	New Jersey	\$12,981	41
X	3	New Mexico	\$7,331	45
X	1	New York	\$12,930	52
X	4	North Carolina	\$6,702	46
	3	North Dakota	\$7,727	49
X	2	Ohio	\$8,963	32
X	4	Oklahoma	\$6,176	2
	3	Oregon*	\$7,619	44
	1	Pennsylvania*	\$9,979	44
	1	Rhode Island	\$9,903	10
X	3	South Carolina	\$7,184	50
	3	South Dakota	\$6,949	54
X	4	Tennessee	\$6,504	37
X	3	Texas	\$7,104	41
	4	Utah	\$5,008	22
	1	Vermont	\$11,128	53
	2	Virginia*	\$8,225	61
X	3	Washington	\$7,243	60
X	2	West Virginia	\$8,475	22
X	1	Wisconsin	\$9,226	71
X	1	Wyoming	\$9,363	58

*State did not use a consistent state criterion referenced fourth grade reading assessment in 2005

APPENDIX B

Appendix B

Table BI: List of State Departments of Education and NAEP websites used to download state assessments and to find contact information for appropriate personnel.

State	Date Accessed Website
Arizona	January 20, 2007 http://www.ade.state.az.us/standards/AIMS/SampleTests/
Arkansas	January 17, 2007 http://arkansased.org/testing/testing.html
California	January 16, 2007 http://www.cde.ca.gov/ta/tg/sr/documents/rtqgr4ela.pdf
Connecticut	February 23, 2007 http://www.csde.state.ct.us/public/cedar/assessment/cmt/index.htm
Florida	January 20, 2007 http://fcats.fldoe.org/fcatsrelease.asp
Georgia	January 20, 2007 http://public.doe.k12.ga.us/ci_testing.aspx?PageReq=CI_TESTING_CRCT
Kentucky	January 20, 2007 http://www.kde.state.ky.us/KDE/Administrative+Resources/Testing+and+Reporting+/District+Support/Link+to+Released+Items/
Louisiana	February 5, 2007 http://www.louisianaschools.net/lde/saa/760.html
Maine	January 31, 2007 http://www.maine.gov/education/mea/04-05ReleasedItems/index.html
Maryland	January 19, 2007 http://www.mdk12.org/mspp/k_8/pr_grade4_reading.html
Massachusetts	January 17, 2007 http://www.doe.mass.edu/mcas/testitems.html
Michigan	January 12, 2007 http://www.michigan.gov/mde/0,1607,7-140-22709_31168_31355-95471-,00.html

State	Date Accessed Website
Mississippi	January 21, 2007 http://www.mde.k12.ms.us/acad/osa/index.html
Montana	December 30, 2006 http://www.opi.state.mt.us/
New Hampshire	February 3, 2007 http://www.ed.state.nh.us/education/doe/organization/curriculum/NECAP/NECAP.htm
New Jersey	January 26, 2007 http://www.nj.gov/education/assessment/es/
New Mexico	January 20, 2007 http://www.ped.state.nm.us/div/acc.assess/accountability/index.html#sab
New York	December 18, 2006 http://www.nysedregents.org/testing/elaei/06exams/home.htm
North Carolina	January 5, 2007 http://www.ncpublicschools.org/accountability/testing/eog/sampleitems/reading4
Ohio	February 4, 2007 http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDetail.aspx?Page=3&TopicRelationID=1070&Content=31225
Oklahoma	February 24, 2007 www.sde.state.ok.us
South Carolina	December 28, 2006 http://ed.sc.gov/agency/offices/assessment/pact/PACTReleaseItems.html
Tennessee	February 23, 2007 http://www.state.tn.us/education/assessment/tsachhome.shtml
Texas	January 20, 2007 http://www.tea.state.tx.us/student.assessment/resources/release/taks/index.html

State	Date Accessed Website
Washington	January 18, 2007 http://www.k12.wa.us/assessment/WASL/testquestions.aspx
West Virginia	December 18, 2006 http://westest.k12.wv.us/filelib.htm
Wisconsin	January 18, 2007 http://www.dpi.state.wi.us/oea/assessmt.html
Wyoming	January 17, 2007 http://www.k12.wy.us/SAA/Paws/index.htm
NAEP	Date Accessed Website
NAEP	January 21, 2007 http://nces.ed.gov/nationsreportcard/itmrls/startsearch.asp