The Cross-sectional and Longitudinal Relationships of Early Childhood School Assessment Policies with Reading Instruction and Reading Achievement: Evidence from Early Childhood Longitudinal Study

by

Haesung Im

A Dissertation Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

Approved September 2015 by the Graduate Supervisory Committee

Kathryn Nakagawa, Chair
Elizabeth Blue Swadener
Marilyn Thompson
Masumi Iida

ARIZONA STATE UNIVERSITY

December 2015
ABSTRACT

The purposes of this study were (1) to examine the direct and indirect effect of school-level testing policies on reading achievement though changes in amount and types of reading instruction, (2) to investigate the reading trajectories moderated by school-level testing policies longitudinally, and (3) to examine the relationship between testing policies and the achievement gap by exploring whether certain student characteristics moderate the relationship between testing policy and reading achievement, using Early Childhood Longitudinal Study Kindergarten (ECLS-K) Cohort of 2010-2011 data. Findings from a multilevel full structural mediation model suggest that school-level frequency of state/local standardized tests had an indirect effect on student reading achievement through changes in both amount and the types of instruction at the school-level (cross-sectional fall kindergarten sample =12,241 children nested in 1,067 kindergarten classes). The findings from a three-level growth models indicated only children of Asian background and children from high socio-economic backgrounds who had frequent standardized tests in kindergarten accelerated in their monthly reading growth, whereas other children (e.g., low SES, non-Asian children) did not show any changes in the rate of the reading growth (longitudinal sample from fall of kindergarten to spring of first grade = 7,392 children nested in 744 kindergartens). The findings from the current study suggest that testing policy is not an effective means to reduce the achievement gap of children from disadvantaged family backgrounds, underperforming children or that children from low socio-economic backgrounds. These children did not seem to benefit from frequent standardized tests longitudinally. Implications for supporting school assessment practices and instruction are discussed.
DEDICATION

To Jesus Christ, my savior
ACKNOWLEDGEMENTS

All praise, honor and glory to Jesus Christ who called me and enabled me to finish this daunting task. I thank Jesus for leading me to examine how testing policy impacts a child's learning process through changes in curriculum and instruction. God provided wonderful mentors and colleagues along the journey who supported me with useful comments and encouragement.

Foremost, I want to express heartfelt gratitude to my mentor, Dr. Kathy Nakagawa for her unwavering support, excellent guidance and high standards. She is an extremely talented scholar. Next, I would like to thank my thesis committee. Dr. Beth Swadener inspired me with her scholarship, community service and passion for justice throughout the program. Dr. Marilyn Thompson, specialized in complex data sets, provided tremendously insightful and thorough critical comments as I sat under her instruction. Special thanks also goes to Dr. Masumi Iida, who helped me to narrow down my research questions during the summer of 2013 using multilevel modeling approach. She made herself available with kindness and great encouragement.

I want to thank other mentors who helped me to grow as an independent scholar. First, I am so grateful to Dr. Elaine Surbeck for connecting the National Board Certification (NBC) process with the Master’s program at Arizona State University enabling me to pursue and research teacher performance assessment as a part of a PhD Program. My special love goes to Susie Baldwin, an accomplished early childhood teacher with great depth of content knowledge and early childhood pedagogy. She allowed me to video-tape and teach science and mathematics lessons to her kindergarteners for over a year. I was privileged to witness exemplary teaching in an
American public kindergarten setting. Also, I cannot thank Dr. Dale Baker enough for her expertise and critical feedback on numerous qualitative researches. I was privileged to sit under the instruction of Dr. Craig Enders, who gave me a deeper appreciation of multilevel and structural equation modeling. I learned so much from Dr. Kyong-Ah Kwon at Georgia State University and Dr. Hyunju Jeon at University of Reno, Nevada about what it takes to complete a published journal in a top tiered peer-reviewed journal. Mathew Valente, a Ph.D. student in quantitative research in psychology in Arizona State University, spent hours discussing with me different methodological challenges. Susie Mauck, a Ph.D. candidate in quantitative studies, Ohio State University helped me to resolve some of the prolonged methodological issues that I had dealt with over the previous three years. Jill McCarroll, a director at National Center for Educational Statistics, helped me to understand the complexity of ECLS-K data structure.

I would also like to express heartfelt gratitude to faculty of Ewha Women's University where I completed my undergraduate and master's programs. The teaching and mentoring of Dr. Eunhye Park, Dr. Yong-Hee Hong, Dr. Heejin Kim, Dr. Jungae Ohm, Dr. Seung Yeon Lee, and Dr. Soon Hwan Kim provided me with strong foundation to complete this daunting task.

Finally, I could not have finished my study without the support and prayers of my husband, Andrew. Andrew was supportive, loving and patient when I was in distress. My four-year old son, Daniel, sparked me with passion for educating young children. And, I also thank my parents, parents-in-law, and grandmother for providing incredible support and prayers along my journey. Ultimately, I thank God for sending me people who showed me endless love and care along the way.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td></td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td></td>
<td>viii</td>
</tr>
<tr>
<td><strong>CHAPTER</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. INTRODUCTION</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Educational Policy Context</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>The Statement of Problem</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Purpose of Study and Research Questions</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>2. PERSPECTIVES ON STANDARDIZED TESTS IN EARLY CHILDHOOD EDUCATION</td>
<td></td>
<td>11</td>
</tr>
<tr>
<td>The Opponents of Standardized Tests</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>3. STANDARDIZED TESTS &amp; READING ACHIEVEMENT</td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>Theoretical and Conceptual Framework</td>
<td></td>
<td>21</td>
</tr>
<tr>
<td>Relationships between High Stakes Testing and Reading Achievement</td>
<td></td>
<td>22</td>
</tr>
<tr>
<td>Longitudinal Relationships between Testing and Student Learning</td>
<td></td>
<td>26</td>
</tr>
<tr>
<td>The Mediating Role of Reading Instruction in the Relationship between Testing and Learning</td>
<td></td>
<td>27</td>
</tr>
<tr>
<td>Present Study</td>
<td></td>
<td>33</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>4. METHODS</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>Data and Analytic Sample</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>Measures</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>Analytic Strategies</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>5. RESULTS</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>Cross Sectional Analyses</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>Longitudinal Analysis of Reading Growth</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>6. DISCUSSION</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>Direct Effect of Testing Policy on Reading Achievement</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>Indirect (mediating) Effects of Testing Policy on Reading Achievement</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>Longitudinal Growth Modeling</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>Limitations and Suggestions for Future Study</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>Implications for Policy and Practices</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>REFERENCES</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>APPENDIX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. TABLES</td>
<td>103</td>
<td></td>
</tr>
<tr>
<td>B. FIGURES</td>
<td>121</td>
<td></td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Description of the Cross-sectional and Longitudinal Samples</td>
<td>104</td>
</tr>
<tr>
<td>2. Descriptive Statistics for State/Local Standardized Tests at Kindergarten</td>
<td>105</td>
</tr>
<tr>
<td>3. Cross-tabulation Between Frequency of Standardized Tests and Reporting Test Scores to Parents</td>
<td>106</td>
</tr>
<tr>
<td>4. Weighted Pearson Product Moment Correlations for Study Variables</td>
<td>107</td>
</tr>
<tr>
<td>5. Weighted Pearson Product Moment Correlations for Study Variables</td>
<td>108</td>
</tr>
<tr>
<td>6. Items Used to Specify Reading Instruction CFA</td>
<td>109</td>
</tr>
<tr>
<td>7. Means, Standard Deviations and Correlations among the Manifest Indicators of Reading Instructional Types</td>
<td>110</td>
</tr>
<tr>
<td>8. Reading Achievement at Spring Kindergarten from Frequency of Standardized Tests</td>
<td>111</td>
</tr>
<tr>
<td>9. Reading Achievement at Spring Kindergarten from Reporting Test Scores to Parents</td>
<td>113</td>
</tr>
<tr>
<td>10. Descriptive Unweighted and Weighted Statistics for Reading achievement</td>
<td>115</td>
</tr>
<tr>
<td>11. Growth Models for Reading Achievement (IRT) from Kindergarten Entry until the Spring of First Grade</td>
<td>116</td>
</tr>
<tr>
<td>12. Random Effects for All Models</td>
<td>120</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>A Conceptual Model Relating Standardized Tests to Student Achievement: .......</td>
<td>122</td>
</tr>
<tr>
<td>2.</td>
<td>Graphical Depiction of the Reading Instruction CFA..................................</td>
<td>123</td>
</tr>
<tr>
<td>3.</td>
<td>Multilevel Full Structural Mediation Models from Kindergarten Frequency of Standardized Tests</td>
<td>124</td>
</tr>
<tr>
<td>4.</td>
<td>Multilevel Full Structural Mediation Models from Kindergarten Reporting Standardized Test Scores to Parents</td>
<td>125</td>
</tr>
<tr>
<td>5.</td>
<td>Examples of Individual Reading Growth over Time ......................................</td>
<td>126</td>
</tr>
<tr>
<td>6.</td>
<td>Linear Slopes for Monthly Reading Growth .............................................</td>
<td>127</td>
</tr>
<tr>
<td>7.</td>
<td>Interaction between Parental income/Occupation/Education and Frequency of Standardized Tests</td>
<td>128</td>
</tr>
<tr>
<td>8.</td>
<td>Interaction between Asian children and Frequency of Standardized Tests........</td>
<td>129</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

Educational Policy Context

The purpose of this study is to examine the relationship between standardized tests at kindergarten and children’s reading achievement concurrently and longitudinally. Also, this study aims to investigate the process through which kindergarten testing policy is related to changes in kindergarten curriculum and instruction, and student learning. After the inception of the No Child Left Behind Act (2001), assessments have played increasingly vital roles in shaping curricular and instructional practice as well as raising student performance (Crocker, 2005).

The common elements of accountability include standards, assessments, and public reporting policies to hold schools and teachers accountable for raising student performance (Goertz & Duffy, 2001). As part of the standards movement, 43 states adopted Common Core State Standards (CCSS) by 2015. CCSS was created to establish clear guidelines about what skills, knowledge, and content children should be able to understand by the end of each grade. Standardized tests are administered in order to evaluate how well schools and districts are performing with the respect to the state standards (McMillan, 2013). Another important part of accountability is public reporting of student performance to student, school, and district. Although there are variations in the way most states and schools report student achievement data, all 50 states require local school districts to publish and report standardized test scores in some ways (Epstein, 2004, 2005; Goertz & Duffy, 2001).
Scholars and experts published a substantial amount of criticism, warnings, and guidelines to inform the direction for current policies that encourage the use of testing without considering young children’s distinct nature and capacities (Hatch, 2002). Recently, a large group of educational researchers petitioned to Congress to stop standardized test-focused policy and school reforms based on their shared concern about the negative effects of standardized tests on teachers and children (Washington Post, 2015). Some scholars believe there should be no high-stakes accountability testing of individual children before the end of third grade (NAEYC, 1988, 2003; Shepard, Kagan, & Wurtz, 1998; Solley, 2007). Young children are difficult to assess with accuracy due to their unique developmental stages and rapid growth (NAEYC, 2003; Shepard et al., 1998). The younger the child, the more demanding it is to get his or her focused attention. Young children are easily distracted and influenced by emotional status or physical needs such as hunger or fatigue (NAEYC, 2003). Children require a specific level of language skills to take standardized tests, and young children are still developing these language skills. Therefore, standardized testing is inappropriate due to young children’s social and cognitive immaturity.

Opponents of early testing based their arguments on developmental theories and pedagogical reasoning whereas the underlying rationale for the current use of standardized assessment came from the framework of efficiency, measurement, and utilization of the standardized test scores. Criticism of standardized tests has dominated the public discourse, and relatively few scholars have voiced their opinions in defense of standardized testing (Wang, Beckett & Brown, 2006; Phelps, 2005; Wang, Beckett, & Brown, 2006). The basic idea of assessment-driven curriculum is to set high standards to
promote rigor of student learning. Measurement-driven assessment focuses on academic skills. Within this framework, assessment is a vital tool for increasing student learning and improving instructional practices, and helps parents to know their children’s test scores. Measurement-driven assessment was perceived as a catalyst to improve instruction. A carefully conceived testing program can yield improvements in students’ basic skills and can influence instruction and curriculum in a positive way (Popham, Cruse, Rankin, Sandifer & Williams, 1985; Wortham, 1995). The underlying assumption is the more students are tested, the more students will be motivated to improve their academic performance.

Despite these theoretical arguments against standardized testing, the use of standardized tests in primary grades has increased rather than decreased. The pressure to collect wide-scale assessment data of young children from the federal to state level has been intensifies rapidly especially after No Child Left Behind (NCLB) (National Research Council, 2008; Scott-Little, Kagan, & Clifford, 2003). To date, the National Assessment of Educational Progress (NAEP), also known as "the Nation's Report Card," assesses representative national samples of students attending public and private elementary schools, junior high schools, and high schools (McLaughlin, 1997). Increasing numbers of states are developing pre-kindergarten standardized assessments for school readiness. Head Start preschoolers take standardized tests as required by the National Reporting System (Epstein, Schweinhart, & DeBruin-Parecki, 2010; Hirsh-Pasek, Golinkoff, Berk, & Singer, 2009). Evidently, using the results of standardized testing will continue to increase as a key instructional instrument in the field of early childhood.
This rapid spread of standardized testing will probably continue to grow due to several factors. First, student performance is used as a criterion for federal and state funding. Federally funded schools are generally required to use measureable student outcomes such as math or reading scores so that they can qualify for federal funding (Meisels, Steele, & Quinn-Leering, 1993). In fact, the pressure to test is frequently associated with receiving funds in a nationwide survey of trends (Jordan & Hackbart, 1999; Shepard, Taylor, & Kagan, 1996). This phenomenon has become widespread with the advent of NCLB and more recently, Race to the Top. Through Race to the Top, $500 million of federal funds has been disseminated to states with preference given to applicants with a strong emphasis on the use of assessment to improve the quality of early childhood education (Ackerman & Coley, 2012). Consequently, a large number of pre-kindergarten stakeholders felt the need to measure young children’s knowledge and skills without full knowledge of the nature of child assessment.

Second, current education reform emphasizes academic achievement as the evidence of success (Shepard & Smith, 1988). While the mechanism for accountability is not only limited to the end results, standardized test scores have rapidly become an all-important tool for measuring educational improvement (Carnoy & Loeb, 2002). Curriculum acceleration, an early jump on content matters, has been a characteristic of ensuring program quality. Since NCLB (2001), skills taught in second grade are now being taught in kindergarten. If students have to read by third grade, children must acquire a certain level of content and standards by first and second grade. Therefore, the academic pressure has been pushed down to the kindergarten.
At issue is the assessment policy of NCLB. Originally designed for secondary grades, it is now impacting the field of early childhood education in considerable ways. For example, at many schools early childhood education curricular focus has shifted from a child-centered priority to the mastery of academic skills, the achievement of predetermined outcomes, and the need for accountability (Goldstein, 2007). In comparison to 20 years ago, current early childhood programs have many more didactic instructions, drills, and worksheets that focus mainly on mere facts (Hirsh-Pasek et al., 2009). The heavy emphasis on academic achievement results in less time spent in free-choice play and out-doors play.

The Statement of Problem

Considering the substantive impact of standardized testing on young children, there is a need for more research on the implications of standardized testing policies in relation to early education. There are three unresolved issues that require examination. First, the link between assessment and student learning has been tenuous due to the lack of empirical studies in early childhood education (NAEYC, 1988, 2003; Solley, 2007). Although there is a body of research evidence in secondary education, current debate about the use of standardized testing during early childhood education largely consists of rhetorical claims and counterclaims. The claim that the use of standardized testing improves student learning remains an empirical question. Scholars and early childhood professionals consistently underscore the necessity of scientific evidence that connects standardized testing and student learning during early childhood. Although there have been some studies that investigated the connection between the strength of schools’
accountability policies and student outcomes during middle childhood and early adolescence (Dee, 2002; Haney, 2000; Lilard & DeCicca, 2001), only a few studies have explored the efficacy of standardized testing during early childhood empirically (Boat, Zorn, Austin, 2005; Hatch & Grieshaber, 2002; Rous, McCormick, Gooden, & Townley, 2007). These studies in early childhood have been limited for generalization due to their small sample sizes, and qualitative research design (Lee & Wong, 2004). Applying empirical findings from secondary education requires caution because the same application of standardized testing may have markedly different impacts on children during early childhood as opposed to middle childhood.

Second, the current body of quantitative research has primarily focused on linear relationships between high-stakes testing and students’ academic outcomes and has not provided any evidence regarding mediating mechanisms (i.e., indirect effects) of high-stakes testing on students’ academic outcomes. They did not investigate a process through which the testing had an impact on the outcomes, especially in early childhood. Curriculum and instruction are an integral part of the cycling process (i.e., assessment-curriculum-learning) that mediates the relationship between high-stakes tests and student learning. The use of standardized testing affects not only the evaluation process but also the overall process of curriculum and instruction. For example, increased high-stakes testing, has led to a change in the focus of the early childhood curriculum from child-centered play-based curriculum to an increasingly academic emphasis of the curriculum (NAEYC, 2003).

Third, the majority of existing studies are inundated with single-level analyses without teasing out school-level effects from student-level effects. For example, previous
studies have failed to tease out the effect of school-level testing policy from student background characteristics (age, initial reading score, race, gender, family SES, home language, and disability) as they tended to employ single-level analyses. Furthermore, the current body of literature has focused primarily on comparing “states” that adopted severe accountability versus “states” that did not have any exposure to standardized tests. Therefore, the findings that focus on state-level analyses may not capture the variations between high-stakes testing at the school level. In response to this need for advanced analytic methods, this study will employ multilevel structural equation mediation modeling. In particular, frequency of standardized tests and reporting to parents are used as important predictors. Therefore, the purpose of this study is to examine the direct and indirect effects of school-level high-stakes testing policies on reading achievement at the end of kindergarten with a multilevel modeling approach. Moreover, it examines the mediating role of reading instruction between kindergarten high-stakes tests and student learning based on curriculum integration theory.

Fourth, a longitudinal study is vital to address questions about whether current standardized testing policies are predictive of long-term learning outcomes, or only result in short-term gains related to teaching to the test. Studies have shown that although the application of accountability might be positively associated with a short-term increase in test scores, the effect is often limited to basic content and skills at the primary grade (Lee, 2008). For instance, teacher-directed didactic instruction during early childhood may be associated with increased cognitive learning outcomes in the short-term, but the effect may fade away longitudinally (Marcon, 2002; Schweinhart & Weikart, 1997). In contrast, a child-centered approach may not produce as much of an increase in test scores in
comparison to didactic instruction in the short term, but may generate greater gain in the long run. Therefore, an investigation of longitudinal effects is needed.

**Purpose of Study and Research Questions**

In relation to these unanswered issues, the purpose of this study was to examine the relationships of school standardized assessment policies during early childhood and reading achievement cross-sectionally and longitudinally. More specifically, it examined: 1) the direct effect of school-level standardized testing policy on kindergarten reading achievement, 2) the indirect effect of school-level standardized testing policies on kindergarten reading achievement through reading instruction, and 3) the long term association of school-level standardized testing policy during early childhood on children’s reading trajectories during the first two years of schooling.

**Research Questions**

In both cross-sectional and longitudinal analysis, I examined two types of school-level standardized testing policies 1) the frequency of state/district standardized tests at school, and 2) whether or not the state/local standardized test scores were reported to parents at kindergarten.

**Cross-sectional Analysis**

In the cross-sectional analyses, the following research questions are posed after controlling for student background characteristics (age in months, initial reading score, race/ethnicity, gender, family SES, home language, and disability) and school-level
environmental factors (e.g., school sector, locale, school’s proportion of non-white students, number of students eligible for free lunch, teachers’ average educational degrees, and teachers’ average years of teaching experiences).

The relationships between frequent standardized tests and reading achievement

1. Is the frequency of standardized tests at kindergarten associated with the amount and the types of reading instruction and reading achievement at kindergarten after controlling for test score reporting policy?

2. Are the amount and the types of reading instruction associated with reading achievement at kindergarten?

3. Do the amount and types of reading instruction mediate the relationship between the frequency of standardized tests at kindergarten and reading achievement at kindergarten after controlling for test score reporting policy?

The relationships between reporting test scores to parents and reading achievement

1. Is the reporting of test scores to parents at kindergarten associated with the amount and types of reading instruction and reading achievement at kindergarten after controlling for frequency of standardized tests?

2. Are the amount and the types of reading instruction associated with reading achievement at kindergarten?

3. Do the amount and types of reading instruction mediate the relationship between the reporting test scores to parents and reading achievement at kindergarten after controlling for frequency of standardized tests?
Longitudinal Analysis of the Effect of Kindergarten Testing Policies

The research questions are posed after controlling for student background characteristics (initial reading score, race/ethnicity, gender, family SES, home language, and disability) and school-level instructional factors (e.g., phonics, integrated language arts, phonics in context) and school-level environmental factors (e.g., school sector, locale, and school’s proportion of non-white students).

1. Do students who have attended schools with different school-level standardized testing policies at kindergarten have different reading achievement growth trajectories from fall of kindergarten to spring of first grade after controlling for child and school-related variables?

2. Do certain student characteristics moderate the relationship between testing policy and reading growth?

Organization of the dissertation

This dissertation is organized as follows. Chapter two discusses perspectives on the standardized testing in early childhood education. Chapter three is the literature review of research evidence on effects of high stakes testing on reading achievement and the mediating role of reading instruction between tests and learning. Methodological design and results are discussed in chapter four and chapter five. Finally, discussions and implications for the study are discussed in chapter six.
CHAPTER 2

PERSPECTIVES ON STANDARDIZED TESTS IN EARLY CHILDHOOD EDUCATION

The controversies of standardized assessment in school accountability reform have been at the center of educational debate especially over the last two decades. The heart of this heated debate stems from conflicting theories, concepts and policies with different visions of schooling and assessments.

The con-side of the argument was from educators, scholars, and teachers. In contrast, the standardized test movement is mainly “initiated by substantial political and policy leadership such as governors, state legislators, state boards of education, and chief state school officers” (Cohen, 1988, p.583). Therefore, while the proponents draw their ideas from effective political system development, cost and benefit analysis, and a market-based approach, the opponents ground their opinions from developmental theories, pedagogical reasons, and philosophical contemplations. Understanding the complexity of the issue is impossible without a solid appreciation of the political, historical, societal, and theoretical contexts. This chapter presents the pro perspectives that promote the use of standardized tests, followed by the con arguments that challenge the very core of the standardized test movement.
The Proponents of Standardized Tests

A standardized test can be defined as any test that is given and scored in a predetermined, standardized manner. Standardized tests are testing instruments that are administered, scored, and interpreted in a uniform manner so that they may be useful for comparisons of same-age children (Martella, 2010). According to Wortham (2012), standardized tests have the following three characteristics: 1) uniform procedures for administration and scoring, 2) allow comparison of student scores by age, grade level, local and national norms, 3) attempt to include material common across most classrooms.

The origin of standardized testing came from creating a mental test for the classification and placement of men (Yerkes, 1919). After President Lyndon Johnson launched his war on poverty, large amounts of federal funding were bestowed to states and schools that reported the program effectiveness via standardized tests (Solley, 2007). In 1965, the Elementary and Secondary Education Act (ESEA) was specifically proposed to support America’s highest poverty schools with a large proportion of disadvantaged students. The report of A Nation at Risk in 1983 demonstrated that the American public education was eroded with a “tide of mediocrity.” It goes on to say, “Average achievement of high school students on most standardized tests is now lower than 26 years ago when Sputnik was launched” (NCEE, 1983, p.8). After the report, an increasing number of schools adapted standardized testing to meet the demand for accountability.

More recently, the 2002 reauthorization of ESEA, renamed the No Child Left Behind (NCLB) Act necessitated all public schools in the United States to test students from third grade to eighth grade. The purpose of this campaign was to “leave no child behind” by improving school accountability through high-stakes testing (Dodge, 2009).
Compelled by the NCLB, literally all schools adopted data collection and reporting systems to be held accountable (Maleyko, 2011). NCLB had greater emphasis on accountability by imposing high-stakes testing on students throughout the nation.

Distinct educational policies and practices always derived from different visions of schooling. The curricular focus on “academic achievement” comes from the idea of global competitiveness and human capital economic theory (Cannella & Swadener, 2006). This market-based paradigm comes from economic theories with cost and benefit analyses (Barnett & Masse, 2007). One dollar invested in high quality early childhood education will yield seven dollars of outcome later in life. The purpose of education is to produce skillful entrepreneurial citizens and workers who will add economic value in the global market (Spring, 2010). The educational goals include: “students’ mastery of subject matter and cognitive skills, as indicated by their performance on standardized tests” (Cohen, 1988, p.586).

Proponents accentuate several benefits of using standardized tests for improving education (Popham et al., 1985; The National Early Childhood Accountability Task Force, 2007; Wortham, 1995). These benefits includes to (1) student learning by raising standards, (2) reducing the achievement gap, (3) benefit of using objective, valid and reliable assessment.

First, proponents of testing believe standardized tests are beneficial for student learning. When there are strong links between learning standards, instruction, and assessment, it provides avenues for focused teaching and evaluation (Crocker, 2005; Martone, & Sireci, 2009; Porter, 1976; Smith, Smith, & DeLisi, 2001). Moreover, when teachers use test scores to make informed instructional decisions, it benefits student
learning. With adequate training and assessment literacy, standardized testing is beneficial for student learning because the curriculum and instruction revolves around curriculum (Crocker, 2005).

Second, proponents of standardized tests concur that assessment data can reduce any achievement gap by providing an extra push to low-performing schools and teachers to be accountable (Popham et al., 1985; Wortham, 1995). Proponents believe that the use of test scores will narrow the achievement gap by monitoring program’s effectiveness in meeting a set of academic standards (Brown & Hattie, 2012; Horton & Bowman, 2002; Merrell; 2012; National Early Childhood Accountability Task Force, 2007; National Research Council, 2008; Phelps, 2005). For teachers in low achieving schools, test scores will be an external mechanism that heightens awareness of the needs of students who may have not progressed at expected rates. For students, test scores will work as a catalyst to motivate underperforming students to study harder to pass certain criteria.

Finally, proponents believe standardized testing is objective and less biased compared to observational assessment tools, which are subject to teachers’ judgment. Standardized tests are administered, scored, and interpreted in a standard manner for comparisons of same-age children (Martella, 2010). They are less biased than the assessor’s subjective judgment, which is easily influenced by students’ race/ethnicity, gender and overall behavior or characteristics (National Early Childhood Accountability Task Force, 2007).
These testing data can be used at multiple levels. At the national level, government can keep track of demographic information, and which programs are effective or not, by disaggregating the information by subgroups (e.g., race/ethnicity, and income level) (National Early Childhood Accountability Task Force, 2007). For parents, test scores may be helpful in choosing a school for their child.

To sum up, drawing from a market-based approach, the proponents of testing argue that use of standardized testing enhances student learning and reduces the achievement gap. However, critics of testing argue whether the same intervention can be applied to all students regardless of their age, ability and distinct needs in each of their critical periods (Miller & Almon, 2009). The opponents of testing contend that standardized tests do not allow for differences in terms of young children’s developmental stages.

**Opponents of Standardized Tests**

To counter the prevailing measurement driven policies and practices, early childhood educators and professionals raised strong opposition to such prevailing emphasis on measurement. In this study, early childhood is defined as birth to eight according to the definition of the National Association of Education for Young Children (NAEYC, 2009), which is one of the largest accredited organizations of early childhood education. Growth stages have been generally divided into three broad stages: early childhood, middle childhood, and adolescence. These developmental stages regarding the nature of young children’s development and growth have been informed by theorists such as Jean Piaget, Lev Vygotsky, Lawrence Kohlberg, and Erik Erikson (Bredekamp, 1987).
Young children are difficult to assess with accuracy before six due to limited abilities and learning styles (Scott-Little et al., 2003; Shepard et al., 1998). Some early childhood professionals recommended that there should be no high-stakes accountability testing of individual children before the end of third grade, and preferably fourth grade due to the unique developmental nature of early childhood (Bredekamp & Rosegrant; 1995; Kamii, 1985; NAEYC, 1990; Shepard et al., 1998). Hence, standardized tests should not be used during early childhood.

Standardized tests are inappropriate for young children because they are developing language, cognition, social-emotional, physical and motor abilities at a rapid pace. Development in young children occurs with huge individual variations in patterns and rates of growth in each of these domains (NAEYC, 1988). Young children learn in ways and rates that are different from older students (Vukelich, Christie, & Enz, 2012). Children require a specific level of language skills to take a standardized test, which children in this age group are barely developing. While standardized testing requires students to communicate through written language by paper and pencil format, young children generally represent their knowledge by showing, talking in natural settings such as during play or daily routines. Also, young children may not understand the meaning of the test or may not understand the importance of doing their best (NAEYC, 2003). The test score measured at one time point in an artificial environment is not adequate to represent the whole dimension of children’s knowledge and capability. Standardized test scores give limited information on young children’s actual knowledge because children could be easily affected by daily emotional status or testing environments.
In contrast to the proponents’ ideas on objectivity, accuracy and efficiency of isolated, standardized testing, the field of early childhood has historically supported ongoing assessment. More specifically, early childhood educators believe teachers’ working knowledge of student learning is much more comprehensive than a single standardized test score. The data for ongoing assessment are collected via multiple resources at multiple times by the teacher. Formative assessments are usually incorporated into classroom routines as an integral part of the instructional process in the context of daily routines to guide and modify instructional strategies (Bennett, 2010; Black & William, 1998a; Dodge, Heroman, Charles & Maiorca, 2004; Garrison & Ehringhaus, 2007). Therefore, teachers could constantly observe children’s skills, prior-knowledge, misconceptions and disposition in interest areas based on the learning objectives of the curriculum, developmental continuum and content standards (Buldu, 2010; Dodges et al., 2004; Puckett & Black, 2000).

Ongoing assessment enables teachers to differentiate instructional strategies for each child with explicit goals. Assessments through multiple sources of information can open the door for enriching curriculum decisions based on a deeper understanding of what and how an individual child learn (Dodge et al., 2004). It provides useful feedback that includes 1) recognition of desired goals, 2) evidence about present position, and 3) a way to close the gap between the two (Sadler, 1989). Descriptive feedback is a crucial component of assessment for learning, because it reduces the discrepancy between the students’ actual level and the reference level (Ramaprasad, 1983).

The strong opposition to the use of standardized testing during early childhood could be explained in relation to the process of developing, implementing, and utilizing
the early childhood large-scale assessment (Scott-Little et al., 2003). First, when designing an assessment, it should reflect the developmental continuum of children’s growth. Assessment of young children should be different from assessment of older students. Assessment design should be differentiated according to young children’s capabilities and their unique developmental stages (NAEYC, 2003; Scott-Little, Kagan, & Clifford, 2003). For example, the use of standardized testing during infants/toddlers’ assessment is not warranted due to the wide-ranging variability and irregularity of early development during this period (National Research Council, 2008). During preschool, assessment for accountability should not be limited to academic disciplines but should comprise all developmental domains. These domains comprise social and emotional development, approaches to learning, language development, cognition and general knowledge.

Second, critics believe that the implementation of standardized testing during early childhood education is extremely challenging because the test results are susceptible to children’s emotional status, culture, command of language, and the context of testing (Bredekamp & Rosegrant, 1995). This is contradictory to proponents’ belief in the efficiency and accuracy of standardized testing. For instance, young children are vulnerable in the context of formal assessment where they often feel stressed and unfamiliar (Black & William, 1998b). Young children are usually not familiar with large-group testing procedures and settings. Hence it must be individualized, with an adult assessor or teacher recording data (National Early Childhood Accountability Task Force, 2007). The process of administering standardized tests among young children is extremely difficult.
Theoretically, each child experiences the assessment similarly given the same instruction under a standard set of administration. However, in practice, there is substantial variability in teachers’ testing behavior as well as in the length of testing sessions and in the physical environment. In an observational study of standardized group testing in ten kindergartens, teachers gave extra help such as cuing correct answers, modified instruction, and applied unique interpretation to the instruction (Wodtke, Harper, Schommer, & Brunelli, 1989). While the teacher experienced interruptions, young children were copying other children’s booklets, verbally cuing other children, and showing inattentive and restless behavior such as humming, singing, or getting out of their seats. Apparently, this study demonstrated that 1) standardized tests are developmentally inappropriate to young children, and 2) it is exceedingly demanding to implement standardized test during early childhood.

Third, at issue is controversy surrounding the area of utilization of the assessment. Standardized testing has been often criticized for merely giving fragmented data on students’ factual knowledge and skills without providing comprehensive information of students’ understanding of concepts, contextual knowledge, and ability for application of theory, students’ strengths. During the kindergarten and primary grades, children may still fail to show their level of confidence under artificial testing conditions, which leads to erroneous conclusions about both the programs and the individual child. Given the fact that young children are developing at a rapid pace, the uses of test scores for making a decision are not appropriate. This is especially more so when the decisions are related to children’s future educational opportunities such as admission or placement.
To sum up, critics’ arguments on developing, implementing, and utilizing standardized testing has been discussed primarily in relation to the distinctive nature of early childhood. It is evident that extensive discussion and criticism against standardized testing has dominated the public discourse over decades (Wang et al., 2006). However, the connection between standardized tests and student learning has been fraught with uncertainties and mixed arguments due to limited scientific evidence especially in the early childhood period.
CHAPTER 3
STANDARDIZED TESTS & READING ACHIEVEMENT

Theoretical and Conceptual Framework

Curriculum integration theory guides this study in relation to how assessment is connected with curriculum and instruction (Applebee, Adler, & Flihan, 2007; Beane, 1997; Soodak & Martin-Kniep, 1994). Assessment is an integral part of the curriculum cycle process of planning, teaching, evaluation, and reflection on teaching and learning (Stringer, Christensen, & Baldwin, 2010). Based on curriculum integration theory, student assessment should also be viewed as an important factor in daily classroom activities because the methods and the use of child assessment are required to align with principles of curriculum and instruction (McAfee & Leong, 2011; Puckett & Black, 2000; Stiggins, Arter, Chappuis, & Chappuis, 2006). Assessment policies and practices impact curriculum and the processes of both teaching and learning. Similarly, the widespread use of standardized tests has influenced the overall process of teaching and learning in kindergarten curriculum.

Within this framework, the decisions about appropriate curriculum and assessment should be based on whether or not use of standardized tests enhance student learning. Therefore, effective curriculum and instruction relies heavily on the types of assessment policies and practices that best promote student learning.

The conceptual framework supporting this reasoning is presented in Figure 1. The conceptual framework illustrates how school-level testing policy impacts student-level reading achievement both directly and indirectly through school-level reading.
instruction. With a multilevel modeling approach, the effect of school-level testing policy on student achievement is examined after accounting for the variance explained by school-level structural factors (e.g., school sector, locale, proportion of non-white students, and proportion of students eligible for free and reduced lunch, teachers’ average educational degree, teachers’ average years of teaching experiences) and student-level characteristics (e.g., child’s initial score, gender, race/ethnicity, home language, age, family socio-economic backgrounds, disability, and initial reading score).

In light of curriculum integration theory, student achievement was modeled with a multilevel 2-2-1 mediation model, which includes a school-level antecedent (X: school-level testing policy), school-level mediator (M: school-level reading instruction), and the student-level outcome variable (Y: student reading achievement) (Zhang, Zyphur, & Preacher, 2009). With this theoretical framework, this study investigated the direct and indirect effects of school-level standardized testing policy on children’s reading achievement.

**Relationships between High Stakes Testing and Reading Achievement**

After the inception of the NBCB (2001), all schools were required to collect comparable test scores via standardized tests and to report test scores to the public in order to be held accountable (Maleyko, 2011).

Therefore, two types of school-level standardized testing policies are of importance: (1) frequency of state and local standardized tests at school-level, and (2) whether or not the school reports state/local standardized test scores to parents.
The school-level frequency of standardized tests reflects the emphasis on raising performance through administering standardized tests. When schools frequently administer state and local standardized tests to their students, it is likely that the schools implemented such tests with the intent of raising student performance. In this sense, school-level frequency of standardized testing could be viewed as a school-level policy.

In the same context, the school’s decision to report standardized test scores to parents can be an important measure of school-level policy that reflects the degree to which a school values accountability to parents. It is likely that schools that opt to report standardized test scores to parents of kindergarteners value holding the school accountable compared to schools that choose not to report the standardized test scores to parents of kindergarteners. Understanding how parents understand test scores during early childhood is especially important because 1) test scores form the basis of a parent’s evaluation of academic progress, 2) parental satisfaction/discontent with local schools depends on test scores, 3) parental attitudes and expectation influence children’s academic success (Barber, Paris, Evans, & Gadsden, 1992).

However, studies indicate that there are almost no formal guidelines for reporting standardized test scores across states, districts, and schools (Barber et al., 1992; Caldwell, 2014). Although how test scores are delivered to parents varies from simple methods such as mail and newsletters to parent-teacher conferences, test scores are typically reported to parents with the simple format of report cards without information on how parents can support their children’s progress (Barber et al., 1992; Caldwell, 2014; Marzano, 2000). Moreover, many teachers experience challenges in how to communicate test scores to parents (DeLuca, & Klinger, 2010; Volante & Fazio, 2007). Thus, parents
are susceptible to misinterpretation of standardized test scores (Afflerbach, 2011).

However, parent-school partnerships and the role of parents are especially important during the early childhood, thus it is possible that reporting standardized test scores to parents may have greater impact on children’s learning trajectories.

Unfortunately, there is no research that has attempted to explore the ways in which school-level standardized testing policies affects student learning during early childhood (NAEYC, 1988, 2003; Solley, 2007). In fact, much of the work has been based on theoretical rather than empirical studies in the field of early childhood (NAEYC, 1988, 2003; Solley, 2007). There is a lack of empirical studies demonstrating how school-level testing policies impact student learning due to limited data at the national level.

Research that links the connection between tests and student learning are fraught with mixed findings. Moreover, there has been a consensus that reading test score trends have been less studied because the reading trajectories have been both inconsistent and far more challenging to interpret than those of mathematics (Lee, 2008; Jacob, 2011). On the one hand, an increasing body of evidence suggests that measurement-driven accountability via standardized testing enhances learning outcomes (Carnoy & Leob, 2002; Hanushek & Raymond, 2004; Kober, Chudowsky, & Chudowsky, 2008). Kober and his colleagues (2008) found that 17 out of 28 states demonstrated moderate-to-large reading gains after applying standardized testing policies. On the other hand, Amrein and Berliner (2002) found that a high-stakes test did not improve student learning as measured in ACT, SAT, NAEP, and AP tests. The analysis of NAEP 4th grade reading scores indicated that only 46% of the states with high-stakes testing had gains in both 1992 and 1998. However, careful examination led to the conclusion that the increased
student achievement was highly associated with the exception of English Language Learners and students with special needs.

Similarly, only a few studies have examined how reporting test scores are related to student learning or curriculum and instruction. However, reporting standardized tests to parents could make a difference on teaching and student learning because teachers may feel pressure to raise student performance. Moreover, parents may be able to use test scores as a basis for understanding their child’s strength and weakness and support student learning at home.

As one of the requirements of NCLB is to communicate test scores to parents (Epstein, 2004, 2005), it is important to investigate how reporting policy may impact student learning. The previous study suggested that states that reported test scores did not have greater reading gain among elementary students (Carnoy & Loeb, 2002; Hanushek & Raymond, 2004). For example, Carnoy and Loeb (2002) suggest that states that merely report test scores did not have any significant improvement in fourth grade reading achievement while states that attached serious consequences such as monetary awards or takeover threats performed significantly better. However, there has been no study that examined the effects of reporting to parents on student achievement during early childhood.

Taken together, the relationship between accountability and frequency of standardized tests at the school-level is unclear. Therefore, there is a need for examining how school-level frequency of standardized tests impacts student’s reading achievement during early childhood. Moreover, the current body of literature from middle childhood focuses primarily on comparing “states” that adopted stringent accountability versus
“states” that did not have any exposure to standardized tests. Accordingly, these studies did not account for the variations across frequency of administering standardized tests or test score reporting practices. For example, the school sector (e.g., private or public) or locale (e.g., urban city, town, or suburb) may affect school-level testing policies. As these studies tended to analyze the effect of state-level testing policy on student achievement without partitioning the variance explained by multiple levels of systems, it is possible that the effect of testing policy could be overestimated due to ignoring the variance explained by school-level factors such as school sector or locale, and demographic characteristics of students (gender, race/ethnicity, home language, initial test score, family SES). Thus, there is a need to study the effect of testing policy on student learning after accounting for the multiple levels of factors that impact student achievement at the school-level.

**Longitudinal Relationships between Testing and Student Learning**

Finally, there are no longitudinal studies that examine the relationship between standardized testing and student learning from early childhood. The NAEP has been considered the best dataset for analyzing state trends. However, it does not have longitudinal data at the student or school-level, which tracks the same students’ learning trajectories over time (Wei, 2008). Although there are longitudinal studies that analyzed states’ reading and math gains from 4th grade to 8th grade, there is no study that kept track of students learning growth moderated by testing policy from kindergarten (Amrein & Berliner, 2002; Braun, 2004; Rosenshine, 2003). Kindergarten is a unique period where children show rapid and sporadic growth. Moreover, young children may not
understand the meaning of the test and could be easily affected by daily emotional status or testing environment. Due to this distinctive characteristics of early childhood, it is possible that the effect of testing during the early childhood may be different from the effects during middle childhood, it is evident that a longitudinal study examining the effect of early childhood standardized testing practices on reading trajectories is needed.

There have been relatively few empirical studies that have employed sound methodology given the limited data available to the public. Thus, questions still remain as to whether higher test scores are predictive of further long-term learning outcomes or merely result in short-term test score gains as a result of teaching to the test. For example, some longitudinal studies indicated that explicit phonics instruction (e.g., code-oriented instruction), is only associated with short-term increases in reading achievement (Kendeou, Broek, White, & Lynch, 2009; Storch & Whitehurst, 2001; Vellutino, 1991). There has been a consensus that reading test score trends have been less studied because the reading trajectories have been both inconsistent and far more challenging to interpret than those of mathematics (Lee, 2008; Dee, & Jacob, 2011). These mixed findings call for further analysis of the connection between standardized testing and student learning outcomes longitudinally.

The Mediating Role of Reading Instruction in the Relationship between Testing and Learning

The current body of research has primarily focused on linear relationships between standardized tests and students’ academic outcomes and has not provided any evidence regarding mediating mechanisms (i.e., indirect effects) between two variables.
Some researchers argue high-stakes testing may narrow the curriculum to only skills needed to pass the standardized test (e.g., Amrein-Beardsley, Berliner, & Rideau, 2010). However, there is no research that has investigated if curriculum and instruction mediate the effects of standardized tests on student outcomes, especially during early childhood. Thus, this study will examine both direct and indirect effects of high-stakes testing on children’s outcomes based on curriculum integration theory. Specifically, the mediating roles of reading instruction will be investigated in two ways: (1) the amount of reading instruction and (2) the types of school-level reading instruction.

**Reading Instructional Time**

Generally, research has indicated that increased reading instructional time is associated with greater gains in reading achievement (Cavanaugh, Kim, Wanzek, & Vaughn, 2004; Chatterji, 2005; Harn, Linan-Thompson, & Roberts, 2008; Simmons, Kame'enui, Harn, Coyne, Stoolmiller, Santoro, & Kaufman, 2007; Sonnenschein, Stapleton, & Benson, 2010). For example, Harn and his colleagues (2008) investigated whether the amount of instructional time affected the reading outcomes for struggling first-grade readers. Students in the less intensive intervention received 30 minutes of instruction for 25 weeks, whereas students in the more intensive intervention received 60 minutes of instruction for 24 weeks. Those who received more instructional time performed significantly better on all reading outcome measures, except comprehension, compared to students who received less instructional time.

Generally, studies suggest high-stakes testing is associated with increased instructional time on tested subjects such as reading and mathematics at the expense of
reducing instruction time in non-tested subject areas such as arts, social studies, music, and physical education (Herman & Abedi, 1994; McAfee & Leong, 1997; McCarty, 2009). For example, McMurrer and Kober’s (2007) study from a survey of 349 school districts during the fifth year of NCLB implementation suggested that 62% of districts reported that they have increased time for reading instruction.

In short, although studies have indicated that high-stake tests are related to increased instructional time on tested subjects during early childhood, few studies have investigated the effect of changed school-level reading instructional time on student achievement. Given that studies generally indicate increased reading instructional time predicts greater reading scores, we can infer that when high-stakes tests are associated with increased reading instructional time, students will learn more. However, the relationship between tests and reading achievement through reading instructional time needs to be explored. Thus, a study is required to connect the relationship between school-level testing policy and reading achievement through changes in the quantity of reading instruction during early childhood.

Types of Reading Instruction

Scholars and educators have different ideas about the types of reading instruction that optimize children’s reading competency (Afflerbach, Pearson, & Paris, 2008; Pearson, 2004). There is a general consensus that the types of instruction make a significant difference in supporting children to read and write.

In this study, reading instruction is categorized in three types by its orientation towards meaning versus decoding skills: (1) phonics, (2) integrated language arts, and (3)
phonics in context. While the phonics instruction has emphasized explicit instruction on decoding skills, integrated language arts highlights the importance of learning language as a whole in a meaningful context. Phonics in context is a new approach that attempts to reconcile these contrasting views of learning to read by teaching phonics in the meaningful context of reading patterned and predictable books. The relationships between each of the instructional approaches and reading achievement are reviewed below.

**Phonics.** One of the distinctive characteristics of phonics instruction is to see the process of reading as a “part to whole” or “bottom-up” process. In other words, acquisition of reading requires young children to decipher and decode written texts. Children need explicit teaching about the form and sounds of letters to decode written language. Phonics is often called a skills-based approach, because it focuses on direct instruction of specific skills such as learning letter names and matching letters to sound.

An increasing number of studies have documented the benefits of explicit phonics instruction particularly on children with low literacy (Juel & Minden-Cupp, 2000; Xue, & Meisels, 2004). For instance, Juel and his colleague (2000) found out that amounts of direct phonics instruction in first grade classrooms led to significant improvements in reading achievement among children below grade level at the start of the school year. Similarly, Morrison and his colleagues (2006) concluded explicit teaching of decoding skills is needed especially for weaker readers. However, a critique of phonics argues that phonics instruction in isolation may not help beginning readers with comprehension while learning letter sound knowledge (Adams, 1990; Krashen, 2002; Weaver, 1990).
**Integrated language arts.** Drawing from *Whole Language* approach philosophy, integrated language arts is a kind of reading instruction that promotes development of reading comprehension through the natural process of reading meaningful texts and through oral communication (Goodman & Goodman, 2009). In other words, reading is learned from whole to part naturally during authentic reading without the need for explicit instruction (Pearson, 2004). Children learn to read during projects with peers. As children engage in a self-selected project, they learn to read by writing, retelling stories, performing plays and skits (Xue, & Meisels, 2004; Sonnenschein et al., 2010). Integrated language arts engages students in reading of authentic materials, writing meaningful compositions and letters, and involves student in integrating their projects with literacy experiences (Pressley, Warton-McDonald, Mistretta-Hampston, & Echevarria, 1998).

Research generally suggests that integrated language arts are associated with greater reading achievement (Krashen, 1999; Maurer, 2010). For example, Krashen (1999) compared “*Whole Language*” and “skills approach” by how much time teachers devote to meaning-oriented reading instruction. The findings suggest that children in classes with more real reading tended to do better on tests of reading comprehension compared to students who had explicit instruction on reading skills. Some scholars found that children with higher language skills benefit more from integrated language arts (Connor et al., 2004; Morrison & Connor, 2002; Morrison et al., 2006; Sonnenschein et al., 2010; Xue & Meisels, 2004). Sonnenschein et al., (2010) kept track of children’s reading growth moderated by different types of instruction (e.g., integrated language arts and phonics instruction) from kindergarten through fifth grade. Findings suggest that children with more advanced phonics skills at kindergarten benefited from an integrated
language arts approach during the kindergarten and first grade, while children who needed intervention did not benefit from an integrated language arts approach.

**Phonics in Context.** Phonics in context is characterized by its attempt to reconcile the debate between phonics and integrated language arts. Going beyond the dichotomy of the phonics versus teaching-for-meaning dilemma, Adam (1990) posits the potential benefits of teaching phonics by using big books and stories with repetition, rhyme. In other words, phonics can be fostered while engaging in purposeful reading of predictable books, stories, rhymes, and songs (Adams, 1990; Juel, 1991, Weaver, 1990). Studies suggest that the types of text matters in learning to read (Hiebert 1999; Jenkins, Peyton, Sanders, & Vadasy, 2004). Specifically, the uses of predictable texts enhance word recognition through “patterned repetitive language that is easy to anticipate and remember” (Johnston, 2000, p. 253) and controlled vocabulary helps children to “focus on print format without being distracted by other cues” (Johnston, 2000, p. 253). Children who are exposed to reading predictable books reported joy of reading (Bridge, Winograd, & Haley, 1983; Grote-Garcia, & Durham, 2013).

Research has indicated children who read predictable, patterned and phonetic texts are likely to develop letter knowledge, phonological awareness, and text decoding skills (Anthony, Lonigan, Burgess, Driscoll, Phillips, & Cantor, 2002; Cunningham, 1990; Grote-Garcia, & Durham, 2013). For example, Cunningham (1990) found that children who learned phonemes within contextualized instruction have significant reading achievement and ability to apply knowledge compared to children in a control group who learned phonemes in isolation /skills from an experimental study. Similarly, Grote-Garcia and Durham (2013) found that children in an experimental group who used
patterned books with repetitive lines and familiar themes learned sight words better than children in a control group because predictable books provided context clues.

Little research has examined the relationship between how testing policy impacts types of reading instruction. Some scholars speculate testing policy would likely increase phonics at the expense of reduced amount of time on integrated language arts as it emphasizes measurable student achievement of certain types of skills and knowledge (Afflerbach, 2011; Pearson, 1996, 2004). Other researchers argue that test construction or the format of test impact the quality of instruction (Au, 2007). For example, standardized tests items that require students to blend sounds may push teachers to have a heavy emphasis on phonics. In contrast, standardized tests that entail samples of narrative or opinion writing would drive teacher instruction towards more time for composing and writing.

Existing literature that connects the relationship between standardized tests and reading instruction are very rare, and tend to be observational/qualitative (Kontovourki, 2009; Zancanella, 1992). Thus, an empirical quantitative study on a large-scale data set is needed to examine how testing policy during early childhood impacts the amount and types of reading instruction.

**Present Study**

Reviews of the literature suggest that there are three gaps that call for future study. First, few empirical studies have been devoted to investigate what actually happens in early childhood curriculum as a consequence of frequent standardized testing. Moreover, these studies have been limited to mostly observational studies in small
classroom settings (Hanes: 2010; Hatch & Grieshaber, 2002; Kallemeyn & DeStefano, 2009; Kontovourki, 2009; Wodtke et al., 1989). Second, the current body of research has primarily focused on linear relationships between high-stakes testing and students’ academic outcomes and has not provided any evidence regarding mediating mechanisms (i.e., indirect effects). Thus, this study will examine the mediating role of reading instruction between high-stakes testing and children’s outcomes. Third, the majority of existing studies are inundated with single-level analyses without teasing out school-level effects from student-level effects. In response to this need for advanced analytic methods, this study will employ multilevel structural equation mediation modeling.

Finally, there are no longitudinal studies that examine the relationship between standardized testing and student learning from early childhood. Given that early childhood is a period where children show rapid and sporadic growth and that studies have suggested that the effect of testing during the early childhood may be different from the effect during middle childhood, it is evident that a longitudinal study examining the effect of early childhood standardized testing practices on reading trajectories is needed.

Building on the previous literature, I hypothesize that testing policy will have a direct effect on reading instructional time. I hypothesize that phonics will be more used than integrated language arts or phonics in context when standardized testing is frequently used. I hypothesize that frequent standardized tests will have a concurrent effect on reading achievement through reading time and phonics near the end of kindergarten. I also hypothesize reporting test scores to parents will have an indirect effect on reading achievement through phonics, phonics in context instruction, which emphasize on decoding skills with different degree. In a similar context, I hypothesize
that standardized tests will be associated with increased reading instructional time. Further, I hypothesize that frequency of tests at kindergarten would not predict student learning at a later grade (e.g., near the end of first grade) given that other studies have found that students who are enrolled in schools with a heavy emphasis on phonics, and phonics in context instruction (decoding skills) did not demonstrate greater reading achievement from first grade (Kendeou, Broek, White, & Lynch, 2009; Sonnenschein et al., 2010; Storch & Whitehurst, 2001; Vellutino, 1991).
CHAPTER 4

METHODS

This study is a secondary data analysis using ECLS-K, 2010-2011 data. For cross-sectional data analyses, a multilevel full structural mediation model was used. For longitudinal data analyses, a three-level growth modeling was conducted using four waves from fall of kindergarten to spring of first grade. The chapter describes sampling, instrumentation, and the mode of analysis to address each research question.

Data and Analytic Sample

The Early Childhood Longitudinal Study- Kindergarten Class of 2010-2011 Cohort (ECLS-K) is an ongoing study sponsored by the National Center for Education Statistics (NCES) within the Institute of Education Sciences (IES) of the U.S. Department of Education. This database is a federally funded longitudinal database based on repeated observations from a nationally representative sample of 20,000 U.S. kindergarten students during the 2010-2011 school years. In addition to the ECLS-K 1998-1999 longitudinal data from kindergarten to eighth grade, the ECLS-K 2010-2011 class data was released in May of 2015. The publicly released data included thus far covers fall kindergarten to spring of first grade (fall of kindergarten, spring of kindergarten, fall of first grade, and spring of first grade) and will continue to keep track of children’s developmental trajectories until eighth grade. Data were collected from parents, teachers, and school administrators. Children in each ECLS-K school were randomly selected from a list of all kindergartners attending that school with stratified multistage sampling.
The sample was limited to (1) children who did not move schools from fall of kindergarten to spring of first grade. In addition, (2) children were excluded when there was a discrepancy between two sets of school-level testing policies, and (3) when cases had missing values or zero on the weight variables. First, children who moved schools from fall of kindergarten to the spring of first grade were excluded because modeling school-level effects on children’s reading growth over time required that only those children who remained in the same school for the duration of the study could be included (Stapleton & Thomas, 2008). Second, kindergarten teachers who responded that they report test scores to parents even when they did not take standardized tests, were excluded. These cases were excluded due to unclear definitions of standardized tests. Third, cases with both missing weights and zero weights were eliminated from analysis because the missing weights indicated that the case was not a participant in the rounds contained in the weight whereas the zero weights indicate that the case participated but did not have the required component (e.g., child assessment) at the round.

The sample included students whose parents reported disability at the fall of kindergarten. As studies have indicated standardized test scores may be inflated and biased with the exclusion of English Language Learners (ELL), current study aimed to examine whether child’s disability status moderate the effect of testing policies (Thompson, DiCerbo, Mahoney, & MacSwan, 2002). The final un-weighted cross-sectional sample consisted of 12,241 children nested in 1,067 kindergarten classrooms in public and private schools nationwide (See Table 1 for further description). There was an average of 11 children from each school.
Measures

**Dependent Variables.** Reading achievement was measured from Item Response Theory (IRT), which estimates latent abilities by evaluating test item responses. For example, IRT score treats the difficulty of each item as information to be incorporated in scaling items based on the probability of a correct/keyed response to an item. Reading IRT scores were measured at four time points: (1) fall of kindergarten, (2) spring of kindergarten, (3) fall of first grade, and (4) spring of first grade. Reading competency comprises holistic measure of reading comprehension, vocabulary knowledge, and children’s basic skills (e.g., print familiarity, letter recognition, beginning and ending sounds, rhyming words, and word recognition). For measuring reading comprehension, children were asked to identify definitions, facts, supporting details, and to make complex inferences within and across texts.

The reading assessment consisted of selected items from Test of Preschool Early Literacy (TOPEL), the Peabody Individual Achievement Test – Revised (PIAT-R), Peabody Picture Vocabulary Test – 3rd Edition (PPVT-III), Preschool Language Assessment Scale (preLas 2000) Form C – Simon Says & Art Show, and the Test of Early Reading Ability – 3rd edition (TERA-3). Additionally, special accommodations were made for English Language Learners. For example, Spanish speakers were assessed with 31 English reading assessment items, which had been translated into Spanish. Across the four waves of data collection weighted mean reading IRT scores were 37.73 (SD = 9.56), 50.06 (SD = 11.45), 56.51 (SD=13.30), and 69.96 (SD = 12.96) for fall of kindergarten, spring of kindergarten, fall of first grade, and spring of first grade, respectively.
Explanatory variables. Two sets of variables that were related to school-level standardized testing policy were used as predictors of analysis: (1) frequency of state/district standardized tests, and (2) reporting state/district standardized tests scores to parents.

Frequency of standardized tests. The school average frequency of state/district standardized tests in the spring of kindergarten was used as a predictor variable. School-level frequency of standardized tests was measured by aggregating the frequency of standardized tests at classroom-level from spring kindergarten teacher questionnaires to school-level at kindergarten. However, the definition of standardized tests was not specifically provided in these teacher questionnaires. For cross-sectional analyses, the frequency scores at the spring of kindergarten were used. The frequency of standardized tests was answered from a teacher questionnaire with a five-point scale. The response categories ranged from “0” being Never to “4” being “three or more times a week”.

Reporting standardized tests scores. The dichotomy variable on reporting test scores to parents was selected from the kindergarten teacher questionnaires. Binary variables, “0” being not reported, and “1” being reported. School-level reporting of standardized test scores to parents was measured by aggregating the classroom-level reporting of standardized test scores to parents at kindergarten.

Two sets of variables related with the instruction were used as mediators: (1) the amount of reading instruction and (2) the type of reading instruction.

The amount of reading instruction. A weekly school-level reading instructional time was measured by spring kindergarten teacher questionnaire. The amount of time spent in language arts instruction was constructed by multiplying the weekly frequency of
instructional time by the duration of instructional time. The school-level instruction time for reading was based on kindergarten teachers’ responses to two items in the spring teacher questionnaire. The frequency of school-level reading instructional time was assessed on a five-point scale: Never, Less than once a week, 1-2 times a week, 3-4 times a week, Daily. The duration of daily instructional time was rated with a four-point scale: 1-30 minutes, 31-60 minutes, 61-90 minutes, and More than 90 minutes. A middle value was assigned to each category in the frequency and the duration measure and then these two scale scores were multiplied (Hong, Corter, Hong, & Pelletier, 2012). The product ranges in value from 0 to 900 minutes, with a mean of about 565.86 minutes per week and a standard deviation of 172 minutes.

**Types of reading instruction.** The types of school-level reading instruction were constructed based on the previous studies by researchers previously, using ECLS-K data (Sonnenschein et al., 2010; Xue & Meisels, 2004). Kindergarten classroom teachers were asked how much time they spent each day teaching each type of reading and language arts skills: no time, half hour or less, about one hour, about two hours, or three hours or more.

The results from a three factor model from a confirmatory factor analysis (CFA) from current study included integrated language arts (6 items), phonics (3 items), and phonics in context (3 items). These items were extracted from kindergarten teacher questionnaires on reading instructional practices (see Table 4). The integrated language arts factor included items about identifying the main idea and parts of a story, communicating complete ideas orally, doing an activity or project related to a book or a story, making predictions based on text, using context cues for comprehension, and
retelling stories. The phonics factor included items about recognizing alphabet letters, matching letters to sounds, and writing one’s own name. The phonics in context factor was comprised of items on using phonetic texts, pattered texts, and controlled vocabulary.

**Control variables.** For control variables, both child and school variables were included. Seven child control variables included: gender, family, socioeconomic backgrounds, race/ethnicity, disability status, home language, age, and initial reading score. Gender was coded 0 being Female and 1 being Male (McCoach, O’Connell, Reis, & Levitt, 2006). SES as a continuous composite variable was constructed by NCES (2015) using five variables: father/male/guardian’s education, mother/female/guardian’s education, father/male/guardian’s occupation, mother/female/guardian’s occupation, and household income. This SES composite was standardized such that it had a mean of 0 and a standard deviation of 1. Since SES composite variables used standardized measures (a mean of 0 and a standard deviation of 1), the range of the SES includes negative values that range from -4.75 to 2.75. Four dummy codes were created for child’s race/ethnicity with White being the reference group (White, Black, Hispanic, Asian, and other race) from the composite race variable (RACE) created at the kindergarten level by National Center for Educational Statistics. A dummy code was created for disability status. Children that were reported to have a disability by parents at kindergarten were coded 1 and children that were not reported to have a disability by parents at kindergarten were coded 0. A dummy variable was created for home language. Children who spoke non-English were coded 1 and children who spoke only English at home were coded 0.
Child’s age in months was subtracted from the time of kindergarten entrance to child’s birthday (e.g., year and months). Finally, child’s initial reading score was assessed at kindergarten fall reading.

School-level control variables included six variables: school locale, sector, percent of students eligible for free and reduced lunch, percent of non-white students, school average teacher education level, and average number of school years of teaching experience as a school teacher. The school locale was coded as a continuous variable, 0 being city, 1 being suburb, 2 being town, and 3 being rural. School sector was controlled because studies have indicated that it is a significant predictor of student achievement (Blank, 2011; Lee & Wong, 2004; Medina, 2009). School sector was coded 0 = public, 1 = private. Percent of non-white students was controlled for because the impact of school policies was moderated by the school racial and ethnic compositions (Hanushek, & Raymond, 2004). The percentage of minority (non-white students) in school was used as a continuous variable and controlled for because studies have shown that schools with different income levels have experienced different levels of influence from accountability policies as well as school sectors (Carbonaro & Covay, 2010; Christensen & Skaerbaek, 2007). Therefore, percent of students eligible for free and reduced lunch was also controlled for. Average teachers’ years of teaching experiences at school was aggregated from teacher level. Average teachers’ education level was also aggregated to school-level. The teachers’ education at class level was rated on a 4-point scale coded as 0=associate’s degree or less, 1=bachelor’s degree, 2=master’s degree, 3=an advanced professional degree beyond a master’s degree.
Finally, average number of school years of teaching experiences was calculated by aggregating the number of school years from each classroom teachers.

**Analytic Strategies**

Two sets of analytic strategies were employed. Cross-sectional mediational models and longitudinal growth models were both conducted to answer different sets of research questions. However, the same longitudinal child sample weight (W4C4P_20) variable was applied in in both cross-sectional and longitudinal sample so that all analyses were drawn from the same sampling weight, and could be generalized to the same population (National Center for Educational Statistics, 2014). A sample weight variable was selected because the sample was based on the child direct assessment questionnaire from fall of kindergarten to spring of first grade. Because of the complex sampling design of the ECLS-K dataset, weight variables were utilized to (1) to make inferences to the population being studied, (2) to adjust for differential sampling rates—e.g., certain groups of children sampled at a higher rate, (3) to adjust for differential nonresponse—e.g., not all parents agreed to be interviewed and whether or not a parent agreed can vary by characteristics of the parent and family (Thomas & Heck, 2001). Without applying the proper sampling weights, parameter estimates would be biased as subpopulations (e.g., Asians and Pacific Islanders) would be oversampled. However, the use of proper weights will correct for different selection probabilities and adjust for effects of nonresponse. Ignoring sampling weight leads to incorrect estimates of population parameters, overestimates of the effects of parameters in predictive models, and inflates Type I error rates.
Cross-sectional Analyses

First, descriptive statistics for school-level frequency of standardized tests, reporting test scores to parents, reading instructional time, reading instructional approach, and reading achievement at spring of kindergarten were obtained.

In the second stage of data analyses, CFA was conducted to investigate the internal structure of the measured items, using Mplus7 (Muthén & Muthén, 2012). Because the measured items in the scale are ordinal, we used a robust mean- and variance adjusted weighted least squares (WLSMV) estimator (Muthén, Du Toit, & Spisic, 1997). Because traditional missing data handling techniques can produce biased parameter estimates or low statistical power (e.g., pairwise and listwise deletion), Full Information Maximum Likelihood (FIML) with Auxiliary variables was used in the present study. Auxiliary variables are extra variables that help in computing the missing cases and boost power (Enders, 2010). Two auxiliary variables were selected that are (a) correlates of the incomplete variables and (b) correlates of missingness to increase power and improve the chances of satisfying Missing at Random (MAR) (Enders, 2010). In this study, Auxiliary variables were included to increase power and to correlate with incomplete variables (e.g., number of books the children had and, how often parents read books to their children).

Finally, a multilevel structural equation model with a mediating variable was conducted by using latent construct (i.e., types of reading instruction from the 3-factor model) as a mediator with Mplus7 (Muthén & Muthén, 2012). All analyses were done after applying the longitudinal weight variable to estimate an unbiased parameter estimate with the nationally representative sample in order to address the current research
questions. A multilevel analysis was conducted because of the multilevel structure of ECLS-K dataset due to the stratified multistage sampling strategies (Thomas & Heck, 2001). The present study had a two-level nested structure: student at Level 1 and school at level 2. The current study modeled the effect of a school-level standardized testing policy on student reading achievement through the effect of school-level reading instruction (see Figure 1 & 2). In order to avoid potential confounding in multilevel mediation, I modeled a Level-2 antecedent ($X$), Level-2 mediator ($M$), and the Level-1 outcome variable ($Y$) (Zhang, Zyphur, & Preacher, 2009). The following equations represent my model, according to the Baron and Kenny (1986)’s three steps. The $a$ paths were defined as the paths from school-level testing policy to school-level mediators (i.e., amount and type of reading instruction). The $b$ paths were defined as the paths from the school-level mediators to student-level reading achievement at spring of kindergarten adjusted for all other mediators and school-level testing policy. The $c'$ path was defined as the path from school-level testing policy to student-level reading achievement at spring of kindergarten adjusted for the effects of the mediators. All paths were adjusted for student-level covariates (gender, age in months, race/ethnicity, family socioeconomic status, initial reading score, home language, and disability status) and school-level covariates (sector, locale, percent of non-white students, percent of free lunch, teachers’ average education, teachers average teaching experiences). In the following model $Test$ included either school-level frequency of standardized tests or school-level test score reporting policy to parents.
Step 1: Estimating $b$ and $c'$ path coefficients

Level 1: $Y_{ij} = \beta_{0j} + \text{Covariates} + r_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01} \text{InstTime}_j + \gamma_{02} \text{Phonics}_j$

$+\gamma_{03} \text{Intergrated Language Arts}_j + \gamma_{04} \text{Phonics in context}_j + \gamma_{05} \text{Test}_j + \text{Covariates} + \mu_{0j}$

Step 2: Estimating A path coefficients

Level 2: $\text{InstTime}_j = \gamma_{10} + \gamma_{11} \text{Test}_j + \mu_{1j}$

Level 2: $\text{Phonics}_j = \gamma_{20} + \gamma_{21} \text{Test}_j + \mu_{2j}$

Level 2: $\text{Integrated language arts}_j = \gamma_{30} + \gamma_{31} \text{Test}_j + \mu_{3j}$

Level 2: $\text{Phonics in context}_j = \gamma_{40} + \gamma_{41} \text{Test}_j + \mu_{4j}$

Additionally, RMediation was used to examine confidence intervals for the indirect effect of interest in the Full Model. Confidence intervals provide a range of possible values for the mediated effect and demonstrate the variability of the effect size (MacKinnon, Fritz, Williams, & Lockwood, 2007; Tofighi & MacKinnon, 2011). In addition, utilizing confidence intervals, compared to normal theory tests of mediation, reduces Type I error, increases statistical power, and produces more accurate confidence limits (MacKinnon, Lockwood, & Williams, 2004). Finally, the RMediation program accounts for the possibility of non-normal distribution of the mediated effect (Tofighi & MacKinnon, 2011).
Longitudinal Analyses

For the longitudinal analyses, a total of four waves were used: fall of kindergarten, spring of kindergarten, fall of first grade, and spring of first grade. Overall there were three general steps using HLM 7 (Raudenbush & Bryk, 2003). First, descriptive analysis of change was analyzed prior to fitting growth curve models to the data. The exploratory analyses included magnitude of change, direction of change, growth form (e.g., linear, quadratic). Individual differences and the rate of change were examined to see whether the shape of growth was linear, quadratic, or piecewise. For instance, students’ SES as a level-2 predictor predicted reading score at level-1. The deviance value of -2 Restricted Log Likelihood (-2RLL/-2LL) was used to compare the relative fits of competing nested models because the difference between two deviances (from the nested models) follows a $\chi^2$ distribution.

Second, the unconditional model was used to test whether there was significant variance among schools by computing the intraclass correlation coefficient (ICC) as well as computing baseline goodness-of-fit for reference (Garson, 2013; Palardy, 2013). ICC measures how related scores are between individuals within the grouping structure. The higher the ICC, the more related students’ scores are to one another’s from the same school. When there are ICCs greater than zero, the assumption of independence of observations in ordinary least squares (OLS) regression is violated. When this assumption is violated, effective sample size is smaller than the observed sample size therefore the standard errors for the OLS estimates in our regression equations will be too small and lead to inflated Type 1 error rates. Therefore, by accounting for the clustering, multilevel modeling adjusts the standard errors in the models.
Finally, a three-level growth model was employed to track students’ reading trajectories from kindergarten to first grade, moderated by school-level standardized testing policies during early childhood. In this study, the school-level frequency of standardized tests at kindergarten and reporting test scores to parents were selected as focal predictors. Restricted Maximum Likelihood (REML) was used to compare nested random effects models that have identical fixed effects, with level-1 being time, level-2 being student, and level-3 being school. Repeated measurement occasions (level-1) are nested within individuals (level-2) and individuals nested within schools (level-3). The level-1 (i.e., within-person) model describes the growth curve for a single individual and the level-2 (i.e., between-person) model describes the average growth curve and the variation in growth between individuals. The level-3 (i.e., between-school) model represents the average growth curve and the variation in growth between schools. At level 3, the β coefficients at level 2 are treated as outcomes to be predicted.

The primary level-1 predictor is a variable that captures the passage of time (measurement wave in this study). As I was interested in predicting final reading achievement, time was centered near the end of first grade. For example, wave 4 being 0, wave 3 being 1, wave 2 being 2, and finally wave 1 being 3. Level-1 variables included the reading IRT scores from the kindergarten to first grade (four waves). Level 2 variables included SES, age at kindergarten entry, gender, and race/ethnicity. Level 2 variables included gender, race/ethnicity, family SES, initial reading score, disability status and home-language.
Level 3 variables were comprised of school locale, sector, proportion of non-white students, and the amount and three types of reading instruction at school level. All variables that did not have a meaningful zero value were group mean centered at all levels of analysis.

Best fitting unconditional model. Before building the model, the normalized weight was entered into the HLM7. Then, the unconditional model was used to test whether there was significant variance among schools by computing the ICC (Garson, 2013; Palardy, 2013). The estimated unconditional model is presented by the following equations.

**Level-1 Model:** \( \text{READING}_{ij} = \pi_{0ij} + \pi_{1ij} (\text{Time}_{ij}) + e_{tij} \)

**Level-2 Model:** \( \pi_{0ij} = \beta_{00j} + r_{0ij} \)
\( \pi_{1ij} = \beta_{10j} \).

**Level-3 Model:** \( \beta_{00j} = \gamma_{000} + u_{00j} \)
\( \beta_{10j} = \gamma_{100} \).

\( \text{READING}_{tij} \) represents the reading score of child \( i \) in school \( j \) at time \( t \). \( \pi_{0ij} \) represents the reading achievement of child \( i \) in school \( j \) at the end of first grade. \( \pi_{1ij} \) represents the rate of reading growth of child \( i \) in school \( j \) between the fall of kindergarten and the spring of first grade. \( e_{tij} \) is the time-specific error of student \( i \) in school \( j \) at time \( t \).

Second, a conditional model (Model A) was estimated with child demographic characteristics and family background as predictors of children’s initial reading achievement and monthly rates of reading growth across three phases. Lists of variable names are described below.
Level-1 Model

\[ \text{READING}_{ij} = \pi_{0ij} + \pi_{1ij} \left( \text{TIME}_{ij} \right) + e_{ij} \]

Level-2 Model

\[
\begin{align*}
\pi_{0ij} &= \beta_{00j} + \beta_{01j} (\text{MALE}_{ij}) + \beta_{02j} (\text{HISP}_{ij}) + \beta_{03j} (\text{BLACK}_{ij}) + \beta_{04j} (\text{ASIAN}_{ij}) \\
&+ \beta_{05j} (\text{OTRACE}_{ij}) + \beta_{06j} (\text{SES}_{ij}) + \beta_{07j} (\text{INITIAL}_{ij}) + \beta_{08j} (\text{DISABL}_{ij}) \\
&+ \beta_{09j} (\text{C1SCR}_{ij}) + r_{0ij} \\
\pi_{1ij} &= \beta_{10j}
\end{align*}
\]

Level-3 Model

\[
\begin{align*}
\beta_{00j} &= \gamma_{000} + u_{00j} \\
\beta_{01j} &= \gamma_{010} \\
\beta_{02j} &= \gamma_{020} \\
\beta_{03j} &= \gamma_{030} \\
\beta_{04j} &= \gamma_{040} \\
\beta_{05j} &= \gamma_{050} \\
\beta_{06j} &= \gamma_{060} \\
\beta_{07j} &= \gamma_{070} \\
\beta_{08j} &= \gamma_{080} \\
\beta_{09j} &= \gamma_{090} \\
\beta_{10j} &= \gamma_{100}
\end{align*}
\]

Where \( \pi_{0ij} \) corresponds to the Level 2 random intercept, \( \beta_{00j} \) is the student-level reading achievement at the end of first grade, \( \beta_{10j} \) is the monthly reading slopes, \( ij \) is individual \( i \) in school \( j \), initial reading represents an indicator variable for initial reading scores, MALE represents Male, HISP represents an indicator variable for Hispanic
(Hispanic=1), Black is African American (Black =1), ASIAN is Asian (Asian =1), OTHRACE is other race (Other race=1), SES is a composite variable for family SES, INITIAL represents child’s initial reading score at the fall of kindergarten, DISABL is children who needed special education (Disability =1), CISCR is whether the child speaks English at home or not ( Non-English at home=1).

Third, a school conditional model (Model B) with kindergarten school-level testing policies as primary predictors of individual child’s reading achievement at the end of first grade was entered.

**Level-1 Model**

\[
\text{READING}_{ij} = \pi_{0ij} + \pi_{1ij} (\text{TIME}_{ij}) + e_{ij}
\]

**Level-2 Model**

\[
\pi_{0ij} = \beta_{00j} + \beta_{01j} (\text{MALE}_{ij}) + \beta_{02j} (\text{HISP}_{ij}) + \beta_{03j} (\text{BLACK}_{ij}) + \beta_{04j} (\text{ASIAN}_{ij})
\]
\[
+ \beta_{05j} (\text{OTHRACE}_{ij}) + \beta_{06j} (\text{SES}_{ij}) + \beta_{07j} (\text{INITIAL}_{ij}) + \beta_{08j} (\text{DISABL}_{ij})
\]
\[
+ \beta_{09j} (\text{CISCR}_{ij}) + r_{0ij}
\]
\[
\pi_{1ij} = \beta_{10j}
\]

**Level-3 Model**

\[
\beta_{00j} = \gamma_{000} + \gamma_{001} (\text{FREQST}_{j}) + \gamma_{002} (\text{REPORT}_{j}) + u_{00j}
\]
\[
\beta_{01j} = \gamma_{010}
\]
\[
\beta_{02j} = \gamma_{020}
\]
\[
\beta_{03j} = \gamma_{030}
\]
\[
\beta_{04j} = \gamma_{040}
\]
\[
\beta_{05j} = \gamma_{050}
\]
\[
\beta_{06j} = \gamma_{060}
\]
\[ \beta_{07j} = \gamma_{070} \]
\[ \beta_{08j} = \gamma_{080} \]
\[ \beta_{09j} = \gamma_{090} \]
\[ \beta_{10j} = \gamma_{100} \]

The focal predictors, FREQST and REPORT, are the frequency of standardized testing at kindergarten and whether the standardized test scores are reported to parents or not at kindergarten. All contextual predictors were grand mean centered at zero to improve interpretability with the exception of dummy codes (Raudenbush & Bryk, 2003).

Third, a school conditional model (Model C) with testing policies at kindergarten was estimated after controlling for school-level weekly reading instructional time and the composite scores of three types of reading instruction at kindergarten.

**Level-1 Model**

\[ \text{READING}_{ij} = \pi_{0ij} + \pi_{1ij} (\text{TIME}_{ij}) + e_{ij} \]

**Level-2 Model**

\[ \pi_{0ij} = \beta_{00j} + \beta_{01j} (\text{MALE}_{ij}) + \beta_{02j} (\text{HISP}_{ij}) + \beta_{03j} (\text{BLACK}_{ij}) + \beta_{04j} (\text{ASIAN}_{ij}) \]
\[ + \beta_{05j} (\text{OTHRACE}_{ij}) + \beta_{06j} (\text{SES}_{ij}) + \beta_{07j} (\text{INITIAL}_{ij}) + \beta_{08j} (\text{DISABL}_{ij}) \]
\[ + \beta_{09j} (\text{C1SCR}_{ij}) + r_{0ij} \]
\[ \pi_{1ij} = \beta_{10j} \]

**Level-3 Model**

\[ \beta_{00j} = \gamma_{000} + \gamma_{001} (\text{FREQST}_{j}) + \gamma_{002} (\text{REPORT}_{j}) + \gamma_{003} (\text{READT}_{j}) + \gamma_{004} \]
\[ (\text{PHONICS}_{j}) + \gamma_{005} (\text{ILA}_{j}) + \gamma_{006} (\text{PHONCNTX}_{j}) + \gamma_{007} (\text{LOCALE}_{j}) + \gamma_{008} \]
\[ (\text{PUPRI}_{j}) + \gamma_{009} (\text{MINOR}_{j}) + u_{00j} \]
Level 3 is between-school model, where \(j\) denotes the school; \(\gamma_{000}\) is student reading achievement at the end of first grade, \(\gamma_{001}\) are the coefficients relating predictors to Level 2 intercept, \(u_{00j}\) is the residual for school \(j\). READT represents school-level weekly reading instructional time, PHONICS represents school-level frequency of phonics reading instruction, whereas ILA is school-level frequency of Integrated language arts and PHONCNTX represents school-level frequency of phonics in context reading instruction. LOCALE is location of school, Private is private school, MINOR is the proportion of non-white students at school.

Fourth, a school conditional model (Model D) with level-2 predictors predicting the monthly reading growth was entered after controlling for school contextual effects.

**Level-1 Model**

\[ \text{READING}_{ij} = \pi_{0ij} + \pi_{1ij}(TIME_{ij}) + e_{ij} \]

**Level-2 Model**

\[ \pi_{0ij} = \beta_{00j} + \beta_{01j}(HISP_{ij}) + \beta_{02j}(BLACK_{ij}) + \beta_{03j}(ASIAN_{ij}) + \beta_{04j}(OTHRACE_{ij}) \]

\[ + \beta_{05j}(SES_{ij}) + \beta_{06j}(INITIAL_{ij}) + \beta_{07j}(DISABL_{ij}) + \beta_{08j}(C1SCR_{ij}) + r_{0ij} \]

\[ \pi_{1ij} = \beta_{10j} + \beta_{11j}(MALE_{ij}) + \beta_{12j}(HISP_{ij}) + \beta_{13j}(BLACK_{ij}) + \beta_{14j}(ASIAN_{ij}) \]

\[ + \beta_{15j}(OTHRACE_{ij}) + \beta_{16j}(SES_{ij}) + \beta_{17j}(INITIAL_{ij}) + \beta_{18j}(DISABL_{ij}) \]

\[ + \beta_{19j}(C1SCR_{ij}) \]

**Level-3 Model**

\[ \beta_{00j} = \gamma_{000} + \gamma_{001}(FREQST_{j}) + \gamma_{002}(REPORT_{j}) + \gamma_{003}(READT_{j}) + \gamma_{004}(PHONICS_{j}) + \gamma_{005}(ILA_{j}) + \gamma_{006}(PHONCNTX_{j}) + \gamma_{007}(LOCALE_{j}) + \gamma_{008}(PUPRI_{j}) + \gamma_{009}(MINOR_{j}) + u_{00j} \]
\[ \beta_{01j} = \gamma_{010} \]
\[ \beta_{02j} = \gamma_{020} \]
\[ \beta_{03j} = \gamma_{030} \]
\[ \beta_{04j} = \gamma_{040} \]
\[ \beta_{05j} = \gamma_{050} \]
\[ \beta_{06j} = \gamma_{060} \]
\[ \beta_{07j} = \gamma_{070} \]
\[ \beta_{08j} = \gamma_{080} \]
\[ \beta_{09j} = \gamma_{090} \]
\[ \beta_{10j} = \gamma_{100} \]

Fifth, a school conditional model (Model E) with level-2 predictors predicting the monthly reading growth and interaction between frequency of standardized tests and level-2 predictors were entered after controlling for school contextual effects.

**Level-1 Model**

\[ \text{READING}_{ij} = \pi_{0ij} + \pi_{1ij}(\text{TIME}_{ij}) + e_{ij} \]

**Level-2 Model**

\[ \pi_{0ij} = \beta_{00j} + \beta_{01j}(\text{MALE}_{ij}) + \beta_{02j}(\text{HISP}_{ij}) + \beta_{03j}(\text{BLACK}_{ij}) + \beta_{04j}(\text{ASIAN}_{ij}) + \beta_{05j}(\text{OTHRACE}_{ij}) + \beta_{06j}(\text{SES}_{ij}) + \beta_{07j}(\text{INITIAL}_{ij}) + \beta_{08j}(\text{DISABL}_{ij}) + r_{0ij} \]

\[ \pi_{1ij} = \beta_{10j} + \beta_{11j}(\text{MALE}_{ij}) + \beta_{12j}(\text{HISP}_{ij}) + \beta_{13j}(\text{BLACK}_{ij}) + \beta_{14j}(\text{ASIAN}_{ij}) + \beta_{15j}(\text{OTHRACE}_{ij}) + \beta_{16j}^{*}(\text{SES}_{ij}) + \beta_{17j}^{*}(\text{INITIAL}_{ij}) + \beta_{18j}^{*}(\text{DISABL}_{ij}) + \beta_{19j}(\text{C1SCR}_{ij}) \]
Level-3 Model

\[ \beta_{00j} = \gamma_{000} + \gamma_{001} (FREQST_j) + \gamma_{002} (REPORT_j) + \gamma_{003} (READT_j) + \gamma_{004} (PHONICS_j) + \gamma_{005} \]

\[ (ILA_j) + \gamma_{006} (PHONCNTX_j) + \gamma_{007} (LOCALE_j) + \gamma_{008} (PUPRI_j) + \gamma_{009} (MINOR_j) \]

\[ + u_{00j} \]

\[ \beta_{01j} = \gamma_{010} \]

\[ \beta_{02j} = \gamma_{020} \]

\[ \beta_{03j} = \gamma_{030} \]

\[ \beta_{04j} = \gamma_{040} \]

\[ \beta_{05j} = \gamma_{050} \]

\[ \beta_{06j} = \gamma_{060} \]

\[ \beta_{07j} = \gamma_{070} \]

\[ \beta_{08j} = \gamma_{080} \]

\[ \beta_{09j} = \gamma_{090} \]

\[ \beta_{10j} = \gamma_{100} + \gamma_{101} (FREQST_j) \]

\[ \beta_{11j} = \gamma_{110} + \gamma_{111} (FREQST_j) \]

\[ \beta_{12j} = \gamma_{120} + \gamma_{121} (FREQST_j) \]

\[ \beta_{13j} = \gamma_{130} + \gamma_{131} (FREQST_j) \]

\[ \beta_{14j} = \gamma_{140} + \gamma_{141} (FREQST_j) \]

\[ \beta_{15j} = \gamma_{150} + \gamma_{151} (FREQST_j) \]

\[ \beta_{16j} = \gamma_{160} + \gamma_{161} (FREQST_j) \]

\[ \beta_{17j} = \gamma_{170} + \gamma_{171} (FREQST_j) \]
\[ \beta_{18j} = \gamma_{180} + \gamma_{181}(FREQST_j) \]

\[ \beta_{19j} = \gamma_{190} + \gamma_{191}(FREQST_j) \]
CHAPTER 5

RESULTS

This chapter is divided into two sections: (1) cross-sectional analyses that focus on the mediational role of curriculum and instruction between kindergarten testing policy and kindergarten reading achievement and (2) longitudinal analyses that center on the effect of testing policy on reading growth over time. In order to reflect the multilevel structure of data due to the multistage probability sampling design, all analyses were conducted within a multilevel modeling framework, accounting for clustering effect at the school-level (Thomas & Heck, 2001).

Sample Description

All descriptive statistics were computed with the longitudinal weight by the use of AM software designed to analyze complex sample survey data (e.g., American Institutes of Research & Cohen, 2005). Standard errors were adjusted with a Taylor series approach. Use of sampling weights accounted for unequal probabilities of selection and nonresponse, and allowed inferences from the nationally representative data.

The final sample for cross-sectional analyses included a total of 12,241 students nested in 1,067 schools. At the student-level, the final weighted sample of children includes 53.6% White children, 11.9% Black children, 23.6% Hispanic children, 5.2% Asian children, and 5.7% children who were either children with more than two races, native American, or Hawaiian (see Table 1). Fifty-one percent of children were male. Twenty percent of children at kindergarten spoke a language other than English at home.
Kindergarten teachers reported that 32.5% of kindergarteners did not take local/state standardized tests whereas a majority of kindergarteners (67.5%) took state and local standardized tests at least once a year (see Table 2). Among those test takers, 20.70% of students took state and district standardized tests more than one or two times a month. The result from cross-tabulation of the weighted sample \((N=11,128)\) indicated that the more frequently that kindergarten classes have standardized tests administered, the more likely the school is to report standardized test scores to parents \(\chi^2(4) = 5852.58, p < 0.001\) (See Table 3).

**Cross Sectional Analyses**

To examine the mediating role of literacy instruction between testing policy and reaching achievement, multilevel structural equation modeling was utilized.

**Bivariate relationships**

Table 4 shows correlations among study variables. Weighted and unweighted correlation coefficients are displayed below and above the diagonal respectively. Children’s reading scores at kindergarten were positively associated with their age. Children who came from higher SES families and had English as a home language were more likely to have higher scores on reading at kindergarten. Children who were Black or Hispanic were more likely to have lower scores on reading whereas Asian children were more likely to have higher scores on reading at kindergarten compared to White children.

Among school-level variables, public schools were more likely to have more students eligible for free lunch and took frequent state/local standardized tests. Public
schools were more likely to administer frequent standardized tests and provide standardized test scores to parents than private schools. Schools that administered frequent standardized tests tended to have teachers with less teaching experience than schools that did not administer standardized tests frequently. Moreover, schools that reported standardized test scores to parents tended to have teachers with more inexperienced teachers and less education.

**Confirmatory Factory Analysis**

Based on the literature on the types of reading instruction from ECLS-K 1998-1999 (Sonnenchein et al., 2010; Xue & Meisels, 2004), the underlying patterns of school-level literacy instruction were identified by Exploratory Factor Analysis (EFA) and CFA using *Mplus* 7 software. All the items were aggregated to the school-level and the analysis was conducted on the final sample of this study. Descriptive statistics were conducted for each item, each subscale, and for the overall score to assess for the normality and missing data.

A two-factor model (e.g., integrated language art versus phonics approach) from previous literature (Sonnenchein et al., Xue & Meisle, 2004) did not fit the data well ($\chi^2(274) = 35627.41, p < .001$. Comparative Fit Index (CFI) = .66, Tucker-Lewis Index (TLI) = .62, Root Mean Square Error of Approximation (RMSEA) = .10, Standardized Root Mean Square Residual (SRMR) = .11). As the two-factor model was based on ECLS-K 1998-1999 data, comparison of CFA between the sample of ECLS-K 1998-1999 and the sample of ECLS-K 2010-2011 was conducted using *Mplus* 7 (Muthén & Muthén, 2012). Results of CFA indicated the factor structures on literacy instruction fitted poorly.
Therefore, an EFA with principal axis factor was employed to identify new constructs from the sample data (i.e., ECLS-K 2010-2011 data). SPSS program was used to conduct Horn’s parallel analysis. Based on a parallel analysis conducted using principal axis factoring (PAF), the mean eigenvalue criterion suggests extracting four factors, whereas O’Connor’s (2000) Minimum Average Partial (MAP) test suggested 3 factors. Therefore, following the MAP test, I extracted three factors from the ECLS-K 2010-2011. I followed the three factor solutions from MAP test because three-factor solution was more consistent with theory and interpretable. These three-factors included phonics, integrated language arts and phonics in context.

Finally, the model fit indices for the three factor model were \( \chi^2 \) (49) = 620.02, \( p < .001 \), CFI = .96, TLI = .94, RMSEA = .03, and SRMR = .03 suggesting that the revised model fit the data well (See figure 2). Studies indicate that the model fit the data well when CFI exceeds .95 (Hu & Bentler, 1999), and when TLI exceeds .90, (Hu & Bentler, 1999), and when RMSEA is less than .08 (Browne & Cudeck, 1993).

Table 6 provides internal consistency estimates of reliability of each of the subscales and the total scale using Cronbach’s alpha. The internal consistency estimate of reliability for the total score in phonics (\( \alpha = .70 \)) was lower than in Integrated language arts (\( \alpha = .81 \)) and phonics in context (\( \alpha = .85 \)). Two residuals of manifest scales which specify an integrated language arts was allowed based on conceptual understanding of each reading instruction (See Green & Hershberger, 2000). Specifically, making prediction based on text is conceptually similar using context cues for comprehension.

Table 7 provides correlation matrices for items in the final sample. Items were listed based on the subscales they belong to. The correlations were moderate between
most of the subscales in each of three factors that represents three types of reading instruction. Table 7 lists the sample size, mean, and the standard deviation for the subscales in reading instruction. The subscale means ranged from 3.04 to 4.80 with the highest scoring subscales being recognizing alphabet and letters, and the lowest scoring subscales being doing an activity or project related to a book or story. The standardized factor loadings ranged from .52 to .85 for phonics, from .44 to .83 for integrated language arts, and from .79 to .84 for phonics in context. Phonics was positively, but weakly correlated with phonics in context, \( r = .17, p < .001 \). Integrated language arts was positively correlated with phonics in context, \( r = .52, p < .001 \), and positively but minimally correlated with phonics, \( r = .14, p < .001 \). The graphic description of the reading instruction CFA is presented in Figure 2.

**Predicting Reading Achievement from Frequency of Standardized Tests**

The model used here is presented in Figure 3. The model fit indices of the model predicting children’s reading achievement at kindergarten from frequency of standardized tests at kindergarten were, \( \chi^2 (201) = 1465.71, p < .001 \), CFI = .95, TLI = .92, RMSEA = .02, and SRMR < .001 (within) and .11 (between), suggesting that the model fit the data well. The ICC, which is the proportion of variance accounted by school-level variables among the total variance in kindergarten reading achievement, was 7.9% in the model.

The results from multilevel full structural mediation analysis shows that frequency of state/local standardized tests was not associated with children’s reading scores near the end of kindergarten after controlling for the covariates including the school-level test
score reporting policy (see Table 8). Differently speaking, children who were enrolled in kindergarten with frequent standardized tests did not have higher reading scores after controlling for covariates. Child control variables included gender, family, socioeconomic backgrounds, race/ethnicity, disability status, home language, age, and initial reading score. School-level control variables comprised school locale, sector, percent of students eligible for free and reduced lunch, percent of non-white students, school average teacher education level, and average number of school years of teaching experience as a school teacher. In the current study model, 71.4% of the variance in children’s reading achievement at kindergarten was explained by child-level variables whereas 18.3% of the variance in children’s reading achievement at kindergarten was explained by and school-level variables.

As presented by unstandardized coefficients, the reading instructional time at school-level was related to children’s reading IRT scores near the end of kindergarten after controlling for school-level testing policy \( (B = .01, SE = .001, p < .05) \). This indicates that children enrolled in schools that spent more time on reading instruction tended to have higher scores near the end of kindergarten than children enrolled in schools that spent less reading instructional time. Frequency of standardized tests at school-level was significantly related to increased time on all types of reading instruction at school-level. For example, when the school took state/local standardized tests more frequently, the school tended to spend more time on phonics \( (B = .060, SE = .02, p < .001) \), on phonics in context \( (B = .16, SE = .03, p < .001) \), and on integrated language arts, \( (B = .13, SE = .03, p < .001) \). Among covariates, proportion of non-white students was related to lower reading scores at kindergarten (see Table 8). The total effect (the
sum of standardized direct and indirect coefficients) of reading instructional time was .09. In other words, a one standard deviation increase in reading instructional time at school level will increase students’ reading achievement by .09 standard deviations on average collapsing across the direct and indirect relations between these two variables. The total effect of phonics instruction was .05, integrated language arts was .08, and phonics in context was .09.

As shown in Table 8, results indicate that there was a mediated effect of frequency of state/local standardized tests on children’s reading scores through reading instructional time ($B = 0.09, SE = .04, p < 0.05$). The result from RMediation program (Tofighi & MacKinnon, 2011) also supported mediated effect of reading instructional time by indicating the 99% confidence limits did not include zero (Lower confidence limit = .01; Upper confidence limit = .19). Thus, kindergarten reading instructional time partially mediated relations between school-level kindergarten frequency of standardized tests and reading achievement at the end of kindergarten.

Additionally, there was a mediated effect of frequency of state/local standardized tests on reading achievement through phonics in context ($B = .10, SE = .05, p < .05$). The results from RMediation was consistent in that 99% confidence interval of phonics in context did not include zero (Lower confidence limit = .01; Upper confidence limit = .20). Therefore, students who are enrolled in kindergarten with frequency of standardized tests had significantly higher reading scores as a result of increased phonics in context reading instruction.

However, no mediated effect was found through Integrated language arts ($B = .07, SE = .05, p > .05$) or through phonics instruction ($B = -.06, SE = .03, p > .05$). However,
when considering the magnitude of coefficients and the large sample size of this study, the mediating effect of frequency was small.

**Predicting Reading Achievement from Reporting Test Scores to Parents**

The model used here is presented in Figure 4. The model fit indices of the model predicting children’s reading achievement at kindergarten were $\chi^2 (201) = 1492.50, p < .001$, CFI = .95, TLI = .92, RMSEA = .02, and SRMR < .001 (within) and .12 (between), suggesting that the model fit the data well. The conditional ICC, which is the proportion of variance in kindergarten reading achievement accounted by school-level variables, was 8.4% in the model. A total of 71.4% of the variance in children’s reading achievement at kindergarten was explained by child-level predictors whereas 18% of the variance in children’s reading achievement at kindergarten was explained by school-level variables.

Table 9 shows that school-level reporting test scores to parents was not associated with children’s reading scores near the end of kindergarten after controlling for the covariates including the school-level frequency of standardized tests. In other words, students who went to kindergarten that reported standardized test scores to parents did not have higher reading scores after controlling for all covariates.

School-level reporting test scores to parents was related to reading instructional time at the school-level ($B = 83.91, SE = 21.19, p < .001$). School-level reporting test scores to parents also was related to phonics in context reading instruction at school-level ($B = .18, SE = .08, p < .05$), but not phonics instruction ($B = .08, SE = .05, p > .05$), or integrated language arts ($B = .07, SE = .06, p > .05$). This indicates that when standardized
test scores are reported to parents, teachers in that school increased only one type of reading instruction (i.e., phonics in context) after controlling for all covariates including test score reporting policy.

The reading instructional time at school-level was associated with children’s reading IRT scores near the end of kindergarten, controlling for reporting test scores to parents ($B = .002, SE = .001, p < .01$). This indicates that children enrolled in schools that spent more reading instructional time tended to have higher scores near the end of kindergarten than children that spent less reading instructional time. The school-level phonics in context was related to higher reading achievement near the end of kindergarten adjusted for kindergarten frequency of standardized tests ($B = .61, SE = .28, p < .05$), while school-level phonics instruction was related to lower reading scores at kindergarten ($B = -1.08, SE = .52, p < .05$). However, integrated language arts was not associated with reading score ($B = .59, SE = .38, p > .05$). Among school-level covariates, private school was related to higher reading scores. However, children in schools with more teaching experience tended to have lower reading scores at kindergarten (see Table 8). The total effect of reading instructional time was .08. In other words, by one standard deviation increases in reading instructional time at school level will increase students’ reading achievement by almost .08 standard deviations on average collapsing across direct and indirect relations between these two variables. The total effect of phonics instruction was .03, integrated language arts was .06, and phonics in context was .07.

As shown in Table 8, results indicate that there was a mediated effect of school-level reporting test scores to parents on children’s reading scores through reading instructional time ($B = 0.18, SE = .08, p < 0.05$). Results from RMediation also indicated
that the 99% confidence limits of mediated effect of reading instructional time did not include zero (Lower confidence limit = .01; Upper confidence limit = .38). Thus, students who were enrolled in those kindergartens that reported standardized test scores to parents had higher reading achievement at the end of kindergarten through increased reading instructional time. However, no mediated effect was found between test reporting policy and reading achievement through phonics \((B = -.09, SE = .08, p > .05)\), or *Integrated language arts* \((B = .04, SE = .05, p > .05)\) or phonics in context \((B = .11, SE = .07, p < .05)\).

**Longitudinal Analysis of Reading Growth**

In order to investigate the longitudinal effect of testing policy at kindergarten over time, multilevel growth modeling was conducted. First, the patterns of growth and missingness and variations between child direct assessments were examined before modeling multilevel growth trajectories. Distributions and the assumptions of multivariate normality were examined using SPSS Version 22.0. The result indicated that the distributions of the variables were approximately normal.

Next, analysis of the patterns of missing data provided evidence that the data were Missing At Random (MAR). The data are MAR if there is a systematic relationship between one or more measured variables and the probability of missing data (Enders, 2010). For instance, if there is a systematic relationship between the missing reading assessment scores and students’ socio-economic background, the data meet the MAR assumption. Modern approaches, in particular missing estimation with the expectation-maximization (EM) algorithm and multiple imputation (MI) or Full Information
Maximum Likelihood (FIML), produce superior estimates when compared with older methods if data are at least MAR (Enders, 2010; Graham, 2012).

Further analyses suggested that a large proportion of data were missing at Level-2 (student) and Level-3 (school). Multiple imputations were used to handle this large proportion of missing data. Multiple imputations result in multiple permutations of the data, each with a different set of plausible replacement values. With careful planning, a single collection of imputed data sets can serve as input for multilevel analyses and increase sample size while boosting power and reducing bias (Enders, 2010). Thus multiple imputations were conducted using NORM 2.03 software (Schafer, 1999). Five data sets were imputed with the EM algorithm to produce pooled results. The final sample included 29,568 observations from 7,392 children nested in 744 kindergartens.

Next, the results from examining the intervals (in months) between children direct assessments indicated that there were large variations in time interval between assessments at each wave. For example, there was an average of six months between each assessment (Max= 9.2 months, Min= 3 months, SD=.97). The time interval between assessments was different because assessment dates ranged from August to December for the fall data collections and from March to June for the spring rounds. These differences in intervals between assessments may affect analyses of gain scores because children assessed in September and June in a given grade have more time to learn skills than children assessed in November and March. Thus, child-specific time was calculated for each individual and applied for all three phases along the non-linear reading growth trajectories.
The results of the unconditional model indicated the 52.32% of variance within
time account for the total variance in reading achievement while 22.85% of variance was
explained by variance among students within same schools and 24.83% of variance was
accounted for by variance among schools. This indicated the three-level multilevel
growth model was needed. For the unconditional growth model for reading achievement
with random intercepts and random slopes, variance in the intercepts was significant ($\tau_{00}
= 50.10, p < .001$) indicating significant amount of variations at the student-level.
Similarly, in the unconditional growth model for reading achievements with random
intercepts and random slopes, variance in the intercepts was significant ($\tau_{00} = 50.10, p
< .001$) indicating significant amount of variations at the school-level. However, the
residual variance for slope was fixed to zero due to low reliability for the best fitting

As Table 11 presents results from model A to E, the conditional student model
(Model A), on average, male children had 0.69 points lower reading scores compared to
female children at near the end of first grade ($\gamma_{010} = -0.69$, SE=0.15, $t (6639) = -4.68
p<.001$). Black children started with lower reading score near the end of first grade
compared to White after controlling for other covariates ($\gamma_{030} = -0.67$, SE=.26, $t (6639) = -
2.56, p<.001$). One unit increase in child’s family socio-economic status was related to
higher reading score near the end of first grade by 1.10 point ($\gamma_{060}=1.10$, SE=.12, $t (6639)
= 8.99, p<.001$). Similarly, children who started with higher reading score at the fall of
kindergarten tended to have a higher average reading score near the end of first grade
($\gamma_{070} = .91$, SE=.01, $t (6639) = 94.04, p<.001$).
Children whose parents reported disability had 1.33 points lower reading score at the end of first grade compared children without parental report on disability ($\gamma_{080} = -1.33$, SE=.18, $t (6639) = -7.24$, $p<.001$).

The findings from Model B suggested that students who went to kindergarten with frequent standardized tests did not have higher reading scores near the end of first grade after controlling for student-level covariates and test score reporting policy ($\gamma_{001} = .36$, SE=.26, $t (741) = 1.37$, $p>.05$). In the same context, students who went to kindergartens with test score reporting policy did not have higher reading scores near the end of first grade after controlling for student-level covariates and school-level frequency of standardized tests ($\gamma_{002} = .20$, SE=.53, $t (741) = .37$, $p>.05$).

According to the results from Model C, students who went to kindergartens with frequency of standardized tests typically did not have higher reading scores near the end of first grade after controlling for student-level and school-level covariates ($\gamma_{001} = 0.38$, SE=.37, $t (741) = 1.40$, $p > 0.05$). Similarly, students who went to kindergartens with test score reporting policy did not have higher reading scores near the end of first grade after controlling for student-level covariates and school-level covariates ($\gamma_{002} = 0.19$, SE=.52, $t (741) = .35$, $p > .05$). Among school-level covariates, only one-type of school-level reading instruction was associated with lower average reading achievement after controlling for student-level and school-level covariates (see Table 9). For example, one unit increase in school-level phonics was associated with an average of .47 points lower reading score near the end of first grade ($\gamma_{004} = -.47$, SE=.19, $t (741) = -2.43$, $p < 0.05$).

Results from Model D shows that on average, reading scores increased by 1.79 points per a month when the child is male, white, average reading score, no disability and
the child is from average family SES and English-speaking family. Male children gained reading scores at significantly slower rate compared to that of the female children ($\gamma_{110} = -0.08, SE = 0.01, t (21,422) = -5.74, p < .001$). Black children showed significantly slower rate of monthly reading growth compared to that of the White children ($\gamma_{120} = -0.05, SE = 0.02, t (21,422) = -2.48, p < .001$). Also, children who started higher at the fall of kindergarten showed slightly slower rate of monthly reading growth from the fall of kindergarten to the spring of first grade ($\gamma_{150} = -0.01, SE = 0.01, t (21,422) = -10.54, p < .001$). Children with disability gained slower reading growth compared to children without disability ($\gamma_{160} = -0.11, SE = 0.01, t (21,422) = -6.29, p < .001$). However, with one unit increase in family SES was related to faster rate of monthly reading growth by 0.1 point per a month ($\gamma_{160} = 0.12, SE = 0.01, t (21,422) = 9.72, p < .001$).

Results from Model E show the cross-level interaction between student-level predictors and school-level frequency of standardized tests after controlling for the student-level and school-level covariates. On average, children showed a monthly reading growth by 1.79 point when there is no standardized test at the school level. On average, frequency of standardized tests did not have any significant effect on the rate of reading growth per a month after controlling for student-level and school-level covariates ($\gamma_{101} = 0.01, SE = 0.01, t (21,422) = .1, p > .05$). With one unit increase in frequency of standardized tests was related to Asian children’s monthly reading growth after controlling for student-level and school-level covariates ($\gamma_{131} = 0.06, SE = 0.03, t (21,422) = 2.03, p < .05$). This suggested that Asian children showed faster rate of monthly reading growth when they were tested frequently at kindergarten after controlling for all covariates. Similarly, children from high SES family background accelerated the rate of
monthly reading growth after controlling for student-level and school-level covariates ($\gamma_{14} = .04, SE = .01, t (21,422) = 2.52, p < .05$). This indicated that children from high SES family accelerated in their reading growth when they went to kindergartens with frequent standardized tests. However, Black, Hispanic, and children with other race/ethnicity did not show any faster rate of monthly reading growth as a result of school-level frequency of standardized tests after controlling for student-level and school-level covariates. Discussions and implications of this study are discussed in chapter six. Table 12 includes random effects for all models.
CHAPTER 6

DISCUSSION

The purpose of this study was to examine the effects of high-stake testing on children’s reading achievement cross-sectionally and longitudinally. In the cross-sectional analyses, this study examined the direct and indirect effects of high-stakes testing on children’s reading scores through school-level reading instructional practices. Longitudinally, this study modeled the reading trajectories moderated by school-level testing policies from near the beginning of kindergarten to near the end of first grade. With advanced analytic methods (e.g., multilevel full structural mediation and three-level growth modeling) from nationally representative data, this study adds to the literature by investigating the longitudinal effect of kindergarten testing policy as well as examining the process through which high-stakes testing impacts reading achievement through changes in the amount and types of reading instruction.

**Direct Effect of Testing Policy on Reading Achievement**

A major finding of this study is that both frequent testing and reporting test scores to parents was not directly associated with children’s reading scores at kindergarten after adjusting for school background variables and individual characteristics. This finding is consistent with previous research indicating that merely reporting test scores was not related to greater reading achievement at fourth grade (Carnoy & Loeb, 2002). However, this finding is inconsistent with the previous studies that found strong accountability (e.g., frequent testing and reporting of scores to parents) is associated with increased student
learning at fourth grade (Kober et al., 2008; Rosenshine, 2003). One possible reason for this discrepancy could be related to the distinctive characteristics of young children at kindergarten. As young children do not understand the purpose of standardized tests, it is likely that testing policy may not motivate kindergarteners to improve test scores or to engage in learning processes compared to fourth grade students. So, frequency of testing and reporting test scores to parents was not directly associated with reading achievement at the end of kindergarten.

Another possible reason may be related to rigor of the analytic strategies especially the use of multiple covariates within the multilevel modeling approach. In contrast to previous studies, the current study teased out the effect of school-level testing policy after controlling for school-level environmental factors (e.g., school sector, locale, school’s proportion of no-white students, number of students eligible for free lunch, teachers’ average educational degrees, and teachers average years of teaching experiences) and student background characteristics (age in month, initial reading score, race/ethnicity, gender, family SES, home language, and disability). In this respect, the effect of accountability might have been overestimated in previous studies.

**Indirect (mediating) Effects of Testing Policy on Reading Achievement**

One of the main findings from this study is that reading instruction time partially mediated the relationships between school-level standardized testing policies (frequency of tests and reporting test scores) and student reading achievements. While previous literature examined mostly linear relationship between standardized tests and student outcome, the present study went a step further and found evidence of a mediating
variable of this effect using sophisticated statistical methods (e.g., multilevel full structural equation mediation model) with a representative dataset.

Findings suggest that school-level frequency of state/local standardized tests had an indirect effect on student reading achievement through changes in both amount and the types of instruction at the school-level. Similarly, the reporting of standardized test scores to parents at the school-level had an indirect impact on increased reading achievement through increased reading instructional time. This finding can be understood in light of what frequency of standardized tests and reporting to parents means both at the school and at the teacher level. At the school-level, when the school took state/local standardized tests more frequently and reported test scores to parents, these schools might have applied more stringent accountability policies compared to schools that test less frequently and opt not to report standardized test scores to parents (Baker et al., 2010). At the teacher level, frequent tests or reporting to parents might have pushed teachers to be more accountable for teaching tested subjects such as reading. Or, it might be that parent involvement caused by the reported test scores stimulated or pushed teachers to devote greater amounts of time to reading instruction.

One of the major findings of this study is that increased reading instructional time mediated the relationship between school-level testing policies and student reading achievement at kindergarten. When the kindergarten took frequent standardized tests and reported test scores to parents, these schools tended to spend more time on reading instruction. Specifically, the more frequently kindergarten classes administered standardized tests, the more time was allocated to reading instruction. Similarly, when the kindergarten classes reported test scores to parents, those kindergarten classes tended to
devote a greater amount of time to reading instruction. These findings are consistent with
previous literature that high-stakes testing is associated with increased instructional time
on tested subjects such as reading and mathematics (Herman & Abedi, 1994; McAfee &
Leong, 1997; McCarty, 2009; McMurrer & Kober, 2007). The fact that weekly reading
instructional time significantly increased as much as 84 minutes a week as a result of
frequent testing signifies young children may have lost instructional time for non-tested
subjects such as arts, social studies, music, and physical education. The more schools
focus on taking standardized tests, the less likely the curriculum will cover non-tested
subjects such as art, science, and critical thinking skills (Herman & Abedi, 1994; Wodtke
et al., 1989). This “teaching to the test” is a more serious issue in early childhood
education where the development of the whole child should be the focus in order to foster
a well-rounded foundation and strong dispositions. Also, the increased reading instruction
time was directly related to reading achievement after controlling for the school-level
testing policies. This finding that more reading instruction time resulted in gains in
students’ reading scores is consistent with previous research findings (Cavanaugh et al.,
2004; Chatterji, 2005; Harn et al., 2008; Simmons et al., 2007; Sonnenschein et al., 2010).

Another finding of this study is that a type of reading instruction (e.g., phonics in
context), partially mediated frequency of standardized testing and students’ reading
achievement. It was found that the more frequently the schools gave standardized tests
the more teachers’ instruction followed phonics in context, which lead to higher student
reading achievement. Although the findings from this study suggest that frequency of
standardized tests was significantly related to changes in all types of reading instruction
at school-level, only one type of reading instruction mediated the relationship between
frequencies of standardized tests and reading achievement. These findings are not consistent with the literature which presupposes the standardized test will increase Phonics but decrease integrated language arts instruction (Pearson, 1996; 2004). Contrary to the hypothesis in the present study, the teachers’ increased emphasis on integrated language arts may be due to the state learning standards. Because state learning standards are the basis of standardized tests (Afflerbach, 2011; McMillan, 2013; McMurrer & Kober, 2007), state learning standards may have a substantial amount of impact on reading instruction at kindergarten (Caldwell, 2014). The Common Core State Standards (CCSS) required kindergarteners since 2010 to (a) demonstrate one-to-one letter-sound correspondences, (b) associate the long and short sounds for the five major vowels, (c) read high-frequency words by sight words, (d) identify sounds of different letters, and (e) read for understanding, retell fiction stories, identify the main topic of an informational text, and retell key details.

As CCSS includes standards that highlight reading for understanding, retelling fiction stories, identifying the main topic of an informational text, teachers may have made increased efforts to help children retell stories, identify main topics with an Integrated language arts reading instruction. In addition, some teachers try to keep the ideal type of instruction for the kindergarteners that they value, namely, developmentally appropriate practice (i.e., whole language approach), in spite of the pressure on schools to emphasize test and decoding skills (Goldstein & Bauml, 2012; McDaniel, Issac, Brooks, & Hatch, 2005; Neuharth-Pritchett, de Atiles, & Park, 2003).

Consistent with the hypothesis, teachers increased phonics instruction in order to align their practice with the CCSS which emphasize decoding skills. However, the
finding that phonics affected reading achievement negatively after controlling for school-level testing policy is not consistent with the preexisting literature. One plausible explanation is that as teachers spent more time in phonics at school, students might have less time for reading books for comprehension. It is important to note that the students’ reading IRT scores measure children’s reading achievement holistically including reading comprehension. So caution should be taken with the interpretation of the results. Given the reading IRT scores were holistic measures of reading comprehension, vocabulary, and word recognition, it is possible that children who were enrolled at schools with more phonics instruction would have lower IRT reading scores. If the measurement of reading achievement focused on decoding skills, there is a possibility that phonics could increase reading scores more.

It is noteworthy that the more frequently schools take standardized tests and report test scores to parents, teachers tend to increase phonics in context as a way of reading instruction in the classroom. Phonics in context might have been promoted due to the standards that require children to read aloud predictably patterned fiction and poetry. Or it may be that teachers may choose to teach phonics in context, because teachers felt torn apart from conflict between phonics and integrated language arts. The finding that students who were taught more with phonics in meaningful and predictable ways with enjoyable books, demonstrated significant improvement in reading scores is consistent with the previous research (Anthony et al., 2002; Cunningham, 1990; Grote-Garcia, & Durham, 2013).
For example, Cunningham (1990) found that learning phonics within contextualized instruction help children to apply knowledge compared to children who learned phonics in isolation. The finding from current study highlights the usefulness of using patterned and predictable books for teaching phonics for young children.

**Longitudinal Growth Modeling**

Findings from this study suggest that children who went to kindergarten with frequency of standardized tests did not have greater average reading score after controlling for student-level and school-level characteristics near the end of first grade. Although there are not enough empirical evidence from previous studies to discuss or support why frequency of standardized tests did not have longitudinal effect on young children’s reading growth, studies on young children’s literacy development suggests that the heavy emphasis on decoding skills may not engender long-term growth in reading (Kendoeou et al., 2009; Storch & Whitehurst, 2001; Vellutino, 1991). However, as the current study kept track of young children’s the reading growth during relatively short-term period (e.g., from kindergarten to the first grade), future study needs to investigate the effect of school-level testing policies with wider span of time points.

Moreover, the findings from three-level growth modeling indicated that school-level frequency of standardized tests had impact on children’s monthly reading growth differently according to children’s race/ethnicity. Specifically, Asian children showed significantly faster rate of monthly reading growth after control for student-level and school-level covariates when they went to kindergarten. This finding is consistent with the previous research that standardized testing has a differential impact across racial,
cultural and age groups (Espinosa, 2005; Dee, 2002; Gordon & Bonilla-Brown, 1996; Gustafason, 2002; Hanushek & Raymond, 2005; Lee & Wong, 2004; McCarthy, 2009). Current study indicates Asian parents might have played a critical role in helping kindergarteners to have greater reading achievement. For example, studies have indicated that both Asian parents spent significantly more time on homework than White American students (Mu, 1997). Also, studies have suggested that Asian parents have high expectation for raising student achievement and interact with their children to promote academic achievement (Hao & Bonstread-Bruns, 1998; Nakagawa, 2008). Accordingly, it is possible that if Asian parents know that their children are tested on a regular basis, Asian parents might have emphasized the mastery of reading at home by providing enriched home literacy environment such as reading books, telling stories, singing songs and taking children to library more often (Kim, Im, & Kwon, in press).

In a similar context, the findings from current study indicated that children from higher family social-economic background accelerated in gaining monthly reading growth from near the beginning of kindergarten to near the end of first grade. This finding is similar to the previous literature that children in high income schools tend to have a greater capability to counter the pressures of external accountability (Carnoy & Loeb, 2002). For example, high-income schools have more resources and greater financial means to purchase extra books, testing materials, and resources for student learning. Similarly, it is possible that parents of high education, high-income and professional job might have supported their young children’s reading growth in various ways when their young children are tested at kindergarten more frequently. Another possibility is that parents of high-income, high education and professional jobs might
have interacted with children with more purposefully to read and write to prepare for tests compared to parents of low-income and less educated parents (Cheadles, 2005; Lareau, 2002, 2003). Lareau (2002) suggest that parents across-diverse socio-economic backgrounds generally have different ways of interacting and engaging in young children’s learning process. For example, parents of high-income and higher education tended to be more comfortable with their children’s teachers and more involved with their children’s schools and details of schooling than were the lower-class parents (Lareau 1989). The finding that only the Asian and children from high SES family benefited from accelerating in their reading growth suggest that frequent testing policy might benefit children’s reading growth when accompanied with parental contribution. 

Taken together, while frequent standardized tests is not detrimental to children, especially subgroup of children who tended to fall behind (e.g., English Language Learns, low-performing child, children with special needs, or Hispanic, Black children), testing policies may not benefit students from underprivileged backgrounds. Therefore, it is difficult to conclude that the uses of standardized tests at kindergarten are effective in reducing the achievement gap.

**Limitations and Suggestions for Future Study**

Until recently, few researchers paid attention to the critical role schools plays in implementing accountability, which impacts overall instruction. This study extended the current knowledge base by examining the relationships of school-level testing policies and student learning after accounting for variance explained by child characteristics and school backgrounds. However, there are several limitations of this study.
First, although standardized tests are generally defined as assessment that is administered, scored, and interpreted in a standard manner, the teacher questionnaire did not specifically provided definition of standardized tests. Thus, definitions of standardized tests were subjective and might vary based on teachers’ judgments.

Second, because school averages of classroom-level reading instructional practices (i.e., the amount and types of reading instruction) was regarded as the reading instruction variables, the variations between teachers within schools were not accounted for in this analysis. Therefore, future studies need to investigate characteristics of individual teachers who have responded to the demands from school-level testing policy differently. For instance, further analyses on teasing out characteristics of teachers who spend greater amounts of time on integrated language arts as opposed to increased phonics instruction in response to the same school-level test reporting policy would add to the current knowledge base for understanding and supporting classroom teachers in the face of increased accountability.

Third, as the reading instructional approach was measured by teacher report in the present study, it is possible that there might be a discrepancy between teachers’ perceived practices and actual practices observed in the classroom (Charlesworth, Hart, Burts, Thomasson, Mosley, & Fleege, 1993; McMullen, 1999). Therefore, future researchers need to conduct more observational studies.

Fourth, the control variables included only external school environment without examining administrators’ and teachers’ pedagogical beliefs or conception of standardized tests. Teachers’ internal belief systems such as pedagogy, content
knowledge, and teachers’ understanding of standardized tests (which may impact the way teacher response) (Brown & Goldstein, 2013; Zancanella, 1992), were not included in this study. Therefore, future studies need to investigate characteristics of schools and teachers that have responded to testing policy differently. For example, further study need to examine the characteristics of school administrators that chose to report test scores versus those who did not would be of critical importance. Similarly, pedagogical beliefs of teachers that provided more phonics or phonics in context as opposed to teachers who spent more time on integrated language arts in response to the same school-level policy would add to the current knowledge base on school-level and teacher-level reaction to accountability policies. Considering the magnitude and small sizes of indirect effects from partial mediation, this study did not fully explain the multiple factors that mediate the relationship between school-level testing policy and student reading achievement. Therefore, other factors such as teachers’ assessment literacy (i.e., understanding of assessment) and content knowledge should be investigated as mediators in future studies.

Fifth, as the current study only investigated how reporting test scores relates to curriculum and instruction at the school-level, future studies need to investigate how school-level test reporting policy impacts parents’ involvement at the school and at home during early childhood. Also, future researchers need to investigate how different test reporting policies (e.g., format or content of report cards) impact parents’ involvement in supporting young children’s academic achievement.

Sixth, in light of curriculum integration theory that explicated the cycling process of assessment-instruction-learning, this study does not include the relationship between state learning standards and assessment. Specifically, future researchers should
investigate how CCSC influences test items in reading, which affects type of instruction especially phonics in context. I suggest future researchers examine if state/local standardized tests measure the early learning standards such as CCSS. Moreover, careful examination is required to test whether the skills and knowledge presented in early learning standards is developmentally appropriate in supporting young children’s learning through valid and reliable standards, assessment and instruction (Martone & Sireci, 2009).

Seventh, future studies need to examine the relationships between standardized testing policy and young children’s social-emotional development (e.g., internalizing, problem-behavior, externalizing, self-control, and approaches to learning) cross-sectionally and longitudinally. The current study only investigated the relationship between school-level testing policy and academic achievement during early childhood. However, as emphasis on academic subjects without considering young children’s developmental stages may impact young children’s social-emotional development negatively (NAEYC, 2003; Shepard et al., 2003). Further research is needed in this area.

Eighth, this study kept track of students only from kindergarten to first grade because of the limitation of publicly released data for analysis. Thus, the present study was limited in tracing the long-term relationships of standardized tests and students’ achievement. Previous studies have suggested that types of instruction may influence children’s reading growth differently depending on their developmental stages (Sonnehchein et al., 2010).
As older and more mature the students are, the more they understand the importance or consequences of raising test performance (Carnoy & Loeb, 2002). Therefore, a longitudinal study that keeps track of students along the wide developmental continuum until fourth grade is necessary.

Finally, examining changes in national trends on the connection between test-instruction-learning is recommended. This is possible by analyzing and comparing ECLS-K 1998-1999 and 2010-2011 data. To the author’s knowledge, there have only been a few studies that examined the relationships of standardized tests and student learning through changes in instruction, using the ECLS-K 1998-1999 data. For example, a study of Im, Kwon, and Jeon (in progress) indicated that the frequency of standardized tests at first grade was associated with poor mathematic achievement at third grade through heavy emphasis on drill instruction at first grade. This finding suggests that there is a possibility that teachers before the NCLB (2001) may have responded to the testing policy differently from those teachers in the academic year of 2010-2011, which is right after the CCSS was introduced in 2010. Thus, comparison of national trends across the 1990s and 2010s would offer timely and significant policy implications.

Implications for Policy and Practices

The findings from current study suggest assessment is an integral part of curriculum and should support student learning. Assessment should support student learning and instruction consistent with the curriculum integration theory. There should be strong links between learning standards, instruction, and assessment. Additionally, assessment tools should be valid and reliable and reflect young children’s developmental
stages (NAEYC, 2011). The most effective literacy instruction is individualized, differentiated instruction tailored to the individual child’s strengths and prior knowledge through ongoing assessment (Morrison, Connor, & Bachman, 2006). Therefore, systematic support for teachers at the school-level is needed to ensure that the cycling process of assessment and instruction supports student learning.

One way to support teachers is to ensure ongoing support for teachers. They need to have assessment literacy which can be defined as the teachers’ ability to interpret and use test scores to enhance student learning (Stiggins, 1991). When teachers understand and recognize different purposes and uses of assessment, assessment literacy will help teachers to make informed decisions under the pressures and demands of accountability from districts and schools. Knowledge on assessment makes a big difference on overall curriculum and in student learning. In contrast, lack of assessment literacy can cause much confusion to teachers and tensions on the academic demands (Brown & Goldstein, 2013; Calkins, Montgomery, & Santman, 1998; DeLuca, & Klinger, 2010; Volante & Fazio, 2007). Therefore, I suggest systematic training and support for early childhood teachers to make them better understand the assessment and communicate test scores to parents in effective ways. Thus, additional support, training, and resources are needed for early childhood teachers to implement the rigorous state standards and use/interpret/communicate results from the state and district standardized tests effectively.

As the findings of the study suggested that frequent standardized tests enhance student learning only through one type of reading instruction (e.g., phonics in context) while not through other types of reading instruction (e.g., phonics & integrated language arts), this study demonstrate that teachers’ choice of effective reading instruction in
response to the demands of testing policy is key to enhanced reading achievement. It is crucial for teachers to understand state standards and to have strong content knowledge, and effective literacy pedagogy in order to enhance students’ reading achievement (McMillan, 2013). In other words, merely implementing more frequent standardized tests to young children would not benefit children’s learning when coupled with ineffective instruction. Thus, supporting early childhood teachers to understand effective pedagogy in teaching reading and literacy is vital.

Additionally, this study highlights the importance of the parent-school partnership in supporting student learning especially during early childhood. In this study, reporting test scores to parents partially mediated student learning through increased reading instruction. However, studies have indicated that many teachers do not know how to communicate to parents, how to interpret test scores, and how to use it to guide instruction (Mertler, & Campbell, 2005; Popham, 2009; Volante, & Fazio, 2007). It is possible that other factors such as parental engagement stimulated by test reporting policy would have enhanced student learning. Moreover, results from longitudinal analysis indicated that children of Asian background and children from high socio-economic family background showed significantly faster reading growth when their school had frequent testing. This finding suggests that frequent standardized tests might be more conducive to student learning when accompanied by parents’ involvement. Hence, systematic support from schools and districts to help parents to be more involved in children’s reading process is recommended.

Kindergarten has a unique place within the U. S. public school system and it has many features that set it apart from other elementary school grade levels (Ehly, 2009;
Goldstein, 2008a; 2008b). To sum up, findings from current study suggest that frequent standardized tests was related to increased reading achievement via changes in types and amount of reading instruction at kindergarten. However, given the small magnitude of effect sizes, the use of standardized tests at kindergarten should be reconsidered after reexamination of its impact on young children’s social-emotional development longitudinally (preferably after third grade).
REFERENCES


International Association of Educators. University of Illinois, Urbana, Retrieved from


Im, H., & Kwon, K., Jeon, H. (In progress). High-stake testing: Accountability for unintended process and mathematic achievements.


98


Spring, J. (2010). *Political agendas for education: From change we can believe in to putting America first*. NY: Routledge


APPENDIX A

TABLES
Table 1

Description of the Cross-sectional and Longitudinal Samples

<table>
<thead>
<tr>
<th>Variable</th>
<th>Cross-sectional Samples</th>
<th>Longitudinal Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unweighted (N=12,241)</td>
<td>Weighted</td>
</tr>
<tr>
<td>Level 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child characteristics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male (%)</td>
<td>51</td>
<td>51</td>
</tr>
<tr>
<td>White (%)</td>
<td>40.3</td>
<td>53.6</td>
</tr>
<tr>
<td>Black (%)</td>
<td>15.3</td>
<td>11.9</td>
</tr>
<tr>
<td>Hispanic (%)</td>
<td>26.4</td>
<td>23.6</td>
</tr>
<tr>
<td>Asian (%)</td>
<td>10.8</td>
<td>5.2</td>
</tr>
<tr>
<td>Other (%)</td>
<td>7.2</td>
<td>5.7</td>
</tr>
<tr>
<td>Initial math score</td>
<td>37.53</td>
<td>37.77</td>
</tr>
<tr>
<td>Non-English household</td>
<td>23.1</td>
<td>20.8</td>
</tr>
<tr>
<td>SES</td>
<td>-.07</td>
<td>-0.07</td>
</tr>
</tbody>
</table>

Note: Weighted by W4C4P_20

*p≤.05, ** p≤.01, *** p≤.001
Table 2

*Descriptive Statistics for State/Local Standardized Tests at Kindergarten*

<table>
<thead>
<tr>
<th>Frequency of State and local standardized Tests,</th>
<th>Unweighted</th>
<th>Weighted</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=10,515</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>19.3%</td>
<td>19.0%</td>
</tr>
<tr>
<td>One or two times a year</td>
<td>52.0%</td>
<td>50.3%</td>
</tr>
<tr>
<td>One or two times a month</td>
<td>23.1%</td>
<td>24.9%</td>
</tr>
<tr>
<td>One or two times a week</td>
<td>2.9%</td>
<td>3.2%</td>
</tr>
<tr>
<td>Three or more times a week</td>
<td>2.7%</td>
<td>2.6%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

**Reporting state and local standardized tests scores to parents,**

N=10,094

<table>
<thead>
<tr>
<th></th>
<th>Unweighted</th>
<th>Weighted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never</td>
<td>25.70%</td>
<td>25.60%</td>
</tr>
<tr>
<td>Yes</td>
<td>74.30%</td>
<td>74.40%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

*Note: Weighted by W4C4P_20*
Table 3

*Cross-tabulation Between Frequency of Standardized Tests and Reporting Test Scores to Parents*

<table>
<thead>
<tr>
<th>Frequency of Standardized Tests</th>
<th>Provide Standardized Test Scores</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unweighted (N=11,128)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Not provided test scores to parents</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>100.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>One or two times a year</td>
<td>15.50%</td>
<td>84.50%</td>
<td>16.60%</td>
<td>83.40%</td>
</tr>
<tr>
<td>One or two times a month</td>
<td>10.30%</td>
<td>89.70%</td>
<td>11.90%</td>
<td>88.10%</td>
</tr>
<tr>
<td>One or two times a week</td>
<td>9.20%</td>
<td>90.80%</td>
<td>10.40%</td>
<td>89.60%</td>
</tr>
<tr>
<td>Three or more times a week</td>
<td>9.60%</td>
<td>90.40%</td>
<td>8.60%</td>
<td>91.40%</td>
</tr>
<tr>
<td>Total</td>
<td>3166</td>
<td>7332</td>
<td>3426</td>
<td>7702</td>
</tr>
</tbody>
</table>

*Note: Weighted by W4C4P_20*

*p≤.05, ** p≤.01, *** p≤.001*
Table 4

Weighted Pearson Product Moment Correlations for Study Variables

<table>
<thead>
<tr>
<th></th>
<th>Reading spring</th>
<th>Reading fall</th>
<th>Home language</th>
<th>AGE</th>
<th>GENDER</th>
<th>SES</th>
<th>HISPANIC</th>
<th>ASIAN</th>
<th>BLACK</th>
<th>WHITE</th>
<th>DISABILITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading spring</td>
<td>-</td>
<td>0.84**</td>
<td>-0.15**</td>
<td>0.14**</td>
<td>-0.08**</td>
<td>0.41**</td>
<td>-0.19**</td>
<td>0.16**</td>
<td>-0.09**</td>
<td>-0.03**</td>
<td>-0.12**</td>
</tr>
<tr>
<td>Reading fall</td>
<td>0.84**</td>
<td>-</td>
<td>-0.15**</td>
<td>0.19**</td>
<td>-0.05**</td>
<td>0.42**</td>
<td>-0.20**</td>
<td>0.16**</td>
<td>-0.07**</td>
<td>-0.04**</td>
<td>-0.08**</td>
</tr>
<tr>
<td>Home language</td>
<td>-0.17**</td>
<td>-0.16**</td>
<td>-</td>
<td>-0.10**</td>
<td>-0.01**</td>
<td>-0.29**</td>
<td>0.46**</td>
<td>0.27**</td>
<td>-0.14**</td>
<td>-0.11**</td>
<td>-0.09**</td>
</tr>
<tr>
<td>AGE</td>
<td>0.15**</td>
<td>0.19**</td>
<td>-0.10**</td>
<td>-</td>
<td>0.06**</td>
<td>0.02**</td>
<td>-0.08**</td>
<td>-0.09**</td>
<td>-0.02**</td>
<td>0.08**</td>
<td>0.09**</td>
</tr>
<tr>
<td>GENDER</td>
<td>-0.08**</td>
<td>-0.05**</td>
<td>0.01**</td>
<td>0.07**</td>
<td>-</td>
<td>-0.01**</td>
<td>-0.01**</td>
<td>-0.03**</td>
<td>0.01**</td>
<td>0.02**</td>
<td>0.13**</td>
</tr>
<tr>
<td>SES</td>
<td>0.40**</td>
<td>0.41**</td>
<td>-0.30**</td>
<td>0.03**</td>
<td>0.01**</td>
<td>-</td>
<td>-0.35**</td>
<td>0.14**</td>
<td>-0.15**</td>
<td>0.04**</td>
<td>-0.01**</td>
</tr>
<tr>
<td>HISPANIC</td>
<td>-0.20**</td>
<td>-0.19**</td>
<td>0.49**</td>
<td>-0.09**</td>
<td>0.01**</td>
<td>-0.33**</td>
<td>-</td>
<td>-0.17**</td>
<td>-0.14**</td>
<td>0.23**</td>
<td>-0.03**</td>
</tr>
<tr>
<td>ASIAN</td>
<td>0.15**</td>
<td>0.16**</td>
<td>0.22**</td>
<td>-0.08**</td>
<td>-0.02**</td>
<td>0.13**</td>
<td>-0.14**</td>
<td>-</td>
<td>-0.14**</td>
<td>-0.50**</td>
<td>-0.08**</td>
</tr>
<tr>
<td>BLACK</td>
<td>-0.08**</td>
<td>-0.06**</td>
<td>-0.12**</td>
<td>-0.02**</td>
<td>0.01**</td>
<td>-0.14**</td>
<td>-0.15**</td>
<td>-0.12**</td>
<td>-</td>
<td>-0.58**</td>
<td>-0.02**</td>
</tr>
<tr>
<td>WHITE</td>
<td>0.00**</td>
<td>-0.01**</td>
<td>-0.05**</td>
<td>0.06**</td>
<td>0.02**</td>
<td>0.07**</td>
<td>0.22**</td>
<td>-0.42**</td>
<td>-0.66**</td>
<td>-</td>
<td>0.06**</td>
</tr>
<tr>
<td>DISABILITY</td>
<td>-0.09**</td>
<td>-0.06**</td>
<td>-0.10**</td>
<td>0.09**</td>
<td>0.13**</td>
<td>-0.01**</td>
<td>-0.04**</td>
<td>-0.07**</td>
<td>-0.03**</td>
<td>0.06**</td>
<td>-</td>
</tr>
</tbody>
</table>

Note: weight is W4C4P_20Weighted coefficients are below the diagonal, whereas unweighted coefficients are above the diagonal.
Table 5

Weighted Pearson Product Moment Correlations for Study Variables (Continued)

<table>
<thead>
<tr>
<th></th>
<th>S2MINOR1</th>
<th>S2PUPRI1</th>
<th>S2STDCO</th>
<th>S2STN</th>
<th>S1HGHSTD</th>
<th>S1YRSTCH</th>
<th>S2LUNCH</th>
<th>X2LOCALE</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2MINOR1</td>
<td>-</td>
<td>-0.14**</td>
<td>0.18**</td>
<td>0.13**</td>
<td>0.02**</td>
<td>-0.11**</td>
<td>0.61**</td>
<td>-0.50**</td>
</tr>
<tr>
<td>S2PUPRI1</td>
<td>-0.14**</td>
<td>-</td>
<td>-0.27**</td>
<td>-0.20**</td>
<td>-0.20**</td>
<td>0.11**</td>
<td>-0.47**</td>
<td>-0.04**</td>
</tr>
<tr>
<td>S2STDCO</td>
<td>0.18**</td>
<td>-0.27**</td>
<td>-</td>
<td>0.43**</td>
<td>-0.03**</td>
<td>-0.07**</td>
<td>0.27**</td>
<td>-0.08**</td>
</tr>
<tr>
<td>S2STN</td>
<td>0.133</td>
<td>-0.20**</td>
<td>0.43**</td>
<td>-</td>
<td>-0.02**</td>
<td>-0.08**</td>
<td>0.20**</td>
<td>-0.06**</td>
</tr>
<tr>
<td>S1HGHSTD</td>
<td>0.014</td>
<td>-0.21**</td>
<td>-0.03**</td>
<td>-0.02**</td>
<td>-</td>
<td>0.14**</td>
<td>0.09**</td>
<td>-0.06**</td>
</tr>
<tr>
<td>S1YRSTCH</td>
<td>-0.11</td>
<td>0.11**</td>
<td>-0.07**</td>
<td>-0.08**</td>
<td>0.14**</td>
<td>-</td>
<td>-0.12**</td>
<td>0.04**</td>
</tr>
<tr>
<td>S2LUNCH</td>
<td>0.613</td>
<td>-0.47**</td>
<td>0.27**</td>
<td>0.20**</td>
<td>0.08**</td>
<td>-0.12**</td>
<td>-</td>
<td>-0.17**</td>
</tr>
<tr>
<td>X2LOCALE</td>
<td>-0.5</td>
<td>-0.04**</td>
<td>-0.07**</td>
<td>-0.06**</td>
<td>-0.06**</td>
<td>0.04**</td>
<td>-0.19**</td>
<td>-</td>
</tr>
</tbody>
</table>

Note: weight is W4C4P_20

Weighted coefficients are below the diagonal, whereas unweighted coefficients are above the diagonal. Values represent results from multi-level analyses accounting for missing values.
Table 6

*Items Used to Specify Reading Instruction CFA*

<table>
<thead>
<tr>
<th>Instructional measure and item</th>
<th>Reliability Cronbach’s Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phonics</strong></td>
<td></td>
</tr>
<tr>
<td>Recognizing alphabet and letters</td>
<td></td>
</tr>
<tr>
<td>Matching letters to sounds</td>
<td></td>
</tr>
<tr>
<td>Writing own name</td>
<td>0.70</td>
</tr>
<tr>
<td><strong>Phonics in Context</strong></td>
<td></td>
</tr>
<tr>
<td>Reading book with phonetic text</td>
<td></td>
</tr>
<tr>
<td>Reading book with patterned text</td>
<td></td>
</tr>
<tr>
<td>Reading book with controlled vocabulary</td>
<td>0.85</td>
</tr>
<tr>
<td><strong>Integrated language arts</strong></td>
<td></td>
</tr>
<tr>
<td>Identifying the main idea and parts of a story</td>
<td>0.81</td>
</tr>
<tr>
<td>Communicating complete ideas orally</td>
<td></td>
</tr>
<tr>
<td>Doing an activity or project related to a book or story</td>
<td></td>
</tr>
<tr>
<td>Making predictions based on text</td>
<td></td>
</tr>
<tr>
<td>Using context cues for comprehension</td>
<td></td>
</tr>
<tr>
<td>Retelling stories</td>
<td></td>
</tr>
</tbody>
</table>

109
Table 7

Means. Standard Deviations and Correlations among the Manifest Indicators of Reading Instructional Types

<table>
<thead>
<tr>
<th></th>
<th>RCGNZ</th>
<th>WRTN</th>
<th>MATC</th>
<th>MAINI</th>
<th>ORALI</th>
<th>DOPR</th>
<th>PREDI</th>
<th>TEXTC</th>
<th>RETEL</th>
<th>CONVO</th>
<th>PATTX</th>
<th>PHOTX</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
<td>4.72</td>
<td>4.51</td>
<td>4.80</td>
<td>3.67</td>
<td>4.46</td>
<td>3.04</td>
<td>4.15</td>
<td>3.98</td>
<td>3.65</td>
<td>3.98</td>
<td>4.16</td>
<td>4.11</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>0.60</td>
<td>0.82</td>
<td>0.44</td>
<td>0.87</td>
<td>0.60</td>
<td>0.91</td>
<td>0.70</td>
<td>0.85</td>
<td>0.77</td>
<td>0.73</td>
<td>0.63</td>
<td>0.71</td>
</tr>
<tr>
<td>RCGNZ</td>
<td>-</td>
<td>0.42</td>
<td>0.60</td>
<td>0.06</td>
<td>0.10</td>
<td>0.06</td>
<td>0.11</td>
<td>0.05</td>
<td>0.03</td>
<td>0.13</td>
<td>0.10</td>
<td>0.12</td>
</tr>
<tr>
<td>WRTN</td>
<td>0.45</td>
<td>-</td>
<td>0.39</td>
<td>0.09</td>
<td>0.07</td>
<td>0.06</td>
<td>0.13</td>
<td>0.05</td>
<td>0.07</td>
<td>0.07</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>MATC</td>
<td>0.60</td>
<td>0.34</td>
<td>-</td>
<td>0.14</td>
<td>0.14</td>
<td>0.11</td>
<td>0.18</td>
<td>0.06</td>
<td>0.17</td>
<td>0.12</td>
<td>0.22</td>
<td>0.11</td>
</tr>
<tr>
<td>MAINI</td>
<td>0.04</td>
<td>0.09</td>
<td>0.14</td>
<td>-</td>
<td>0.42</td>
<td>0.38</td>
<td>0.34</td>
<td>0.29</td>
<td>0.36</td>
<td>0.33</td>
<td>0.35</td>
<td>0.30</td>
</tr>
<tr>
<td>ORALI</td>
<td>0.06</td>
<td>0.05</td>
<td>0.13</td>
<td>0.40</td>
<td>-</td>
<td>0.26</td>
<td>0.51</td>
<td>0.51</td>
<td>0.37</td>
<td>0.24</td>
<td>0.29</td>
<td>0.22</td>
</tr>
<tr>
<td>DOPR</td>
<td>0.04</td>
<td>0.05</td>
<td>0.35</td>
<td>0.23</td>
<td>-</td>
<td>0.34</td>
<td>0.29</td>
<td>0.36</td>
<td>0.20</td>
<td>0.20</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>PREDI</td>
<td>0.10</td>
<td>0.11</td>
<td>0.18</td>
<td>0.63</td>
<td>0.45</td>
<td>0.33</td>
<td>-</td>
<td>0.69</td>
<td>0.47</td>
<td>0.32</td>
<td>0.36</td>
<td>0.29</td>
</tr>
<tr>
<td>TEXTC</td>
<td>0.03</td>
<td>0.06</td>
<td>0.08</td>
<td>0.60</td>
<td>0.48</td>
<td>0.31</td>
<td>0.67</td>
<td>-</td>
<td>0.42</td>
<td>0.35</td>
<td>0.33</td>
<td>0.32</td>
</tr>
<tr>
<td>RETEL</td>
<td>0.03</td>
<td>0.06</td>
<td>0.19</td>
<td>0.51</td>
<td>0.36</td>
<td>0.34</td>
<td>0.38</td>
<td>0.41</td>
<td>-</td>
<td>0.26</td>
<td>0.34</td>
<td>0.29</td>
</tr>
<tr>
<td>CONVO</td>
<td>0.10</td>
<td>0.06</td>
<td>0.10</td>
<td>0.32</td>
<td>0.25</td>
<td>0.19</td>
<td>0.31</td>
<td>0.36</td>
<td>0.24</td>
<td>-</td>
<td>0.64</td>
<td>0.68</td>
</tr>
<tr>
<td>PATTX</td>
<td>0.07</td>
<td>0.08</td>
<td>0.18</td>
<td>0.33</td>
<td>0.31</td>
<td>0.20</td>
<td>0.34</td>
<td>0.33</td>
<td>0.30</td>
<td>0.63</td>
<td>-</td>
<td>0.68</td>
</tr>
<tr>
<td>PHOTX</td>
<td>0.12</td>
<td>0.10</td>
<td>0.12</td>
<td>0.31</td>
<td>0.27</td>
<td>0.20</td>
<td>0.31</td>
<td>0.32</td>
<td>0.28</td>
<td>0.67</td>
<td>0.69</td>
<td>-</td>
</tr>
</tbody>
</table>

*Note:* Weighted coefficients are below the diagonal. RCGNZ= Recognizing alphabet and letters; WRTN= Writing own name; MATC= Matching letters to sounds; MAINI= Identifying the main idea and parts of a story; ORALI= Communicating complete ideas orally; DOPR= Doing an activity or project related to a book or story; PREDI= Making predictions based on text; TEXTC= Using context cues for comprehension; RETEL= Retelling stories; CONVO= Controlled vocabulary; PATTX = Patterned text, PHOTX = Phonetic text.
Table 8

Reading Achievement at Spring Kindergarten from Frequency of Standardized Tests (Unstandardized Coefficients)

<table>
<thead>
<tr>
<th>Predicting Reading Achievement at Spring Kindergarten Predictor</th>
<th>Phonics</th>
<th>Integrated language arts</th>
<th>Phonics in context</th>
<th>Reading (Spring)Kindergarten</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child level</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial Reading Score</td>
<td>0.94 (0.01)**</td>
<td>0.90 (0.13)**</td>
<td>-0.69 (0.14)**</td>
<td>0.94 (0.01)*****</td>
</tr>
<tr>
<td>Sex (female)</td>
<td>-0.69 (0.14)**</td>
<td>-0.45 (0.24)</td>
<td>0.61 (0.39)</td>
<td>-0.68 (0.34) *</td>
</tr>
<tr>
<td>SES</td>
<td>0.90 (0.13)**</td>
<td>0.61 (0.39)</td>
<td>-0.45 (0.24)</td>
<td>0.90 (0.13)*****</td>
</tr>
<tr>
<td>Hispanic</td>
<td>-0.69 (0.14)**</td>
<td>0.61 (0.39)</td>
<td>-0.45 (0.24)</td>
<td>-0.68 (0.34) *</td>
</tr>
<tr>
<td>Asian</td>
<td>0.61 (0.39)</td>
<td>-0.45 (0.24)</td>
<td>0.61 (0.39)</td>
<td>0.61 (0.39)</td>
</tr>
<tr>
<td>Black</td>
<td>-0.68 (0.34) *</td>
<td>0.61 (0.39)</td>
<td>-0.45 (0.24)</td>
<td>-0.68 (0.34) *</td>
</tr>
<tr>
<td>White</td>
<td>0.07 (0.34)</td>
<td>0.61 (0.39)</td>
<td>-0.45 (0.24)</td>
<td>0.07 (0.34)</td>
</tr>
<tr>
<td>English at home</td>
<td>-0.64 (0.25)*</td>
<td>-0.64 (0.25)*</td>
<td>-0.64 (0.25)*</td>
<td>-0.64 (0.25)*</td>
</tr>
<tr>
<td>Student Age in months</td>
<td>-0.03 (0.02)</td>
<td>-0.03 (0.02)</td>
<td>-0.03 (0.02)</td>
<td>-0.03 (0.02)</td>
</tr>
<tr>
<td>Child with disability</td>
<td>-1.22 (0.21)**</td>
<td>-1.22 (0.21)**</td>
<td>-1.22 (0.21)**</td>
<td>-1.22 (0.21)*****</td>
</tr>
</tbody>
</table>

Note: *p≤.05, ** p≤.01, *** p≤.001
Weighted by W4C4P_20
### Table 8

**Reading Achievement at Spring Kindergarten (Continued)**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>School-level</th>
<th>( \text{Reading instructional time} )</th>
<th>( \text{Phonics} )</th>
<th>( \text{Integrated language arts} )</th>
<th>( \text{Phonics in context} )</th>
<th>( \text{Reading achievement} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of minority student</td>
<td>-0.02 (0.01)**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% of free lunch</td>
<td>0.02 (0.01)**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Private school</td>
<td>1.15 (0.52)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reporting ST</td>
<td>0.32 (0.29)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency of ST</td>
<td>42.46 (9.73)**</td>
<td>0.06 (0.02)**</td>
<td>0.13 (0.03)**</td>
<td>0.16 (0.03)**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher education</td>
<td>-0.46 (0.34)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teaching experiences</td>
<td>-0.06 (0.02)**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location of school</td>
<td>0.19 (0.14)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading instruct time</td>
<td>0.01 (0.01)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phonics</td>
<td>-1.09 (0.52)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Integrated language</td>
<td>0.58 (0.28)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phonics in context</td>
<td>0.61 (0.28)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual variance</td>
<td>32540.40 (1637.85)</td>
<td>.14 (.04)</td>
<td>.25 (.04)</td>
<td>.38 (.04)</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Note: *\( p \leq 0.05 \), ** \( p \leq 0.01 \), *** \( p \leq 0.001 \)

Weighted by W4C4P_20
Table 9  
*Reading Achievement at Spring Kindergarten from Reporting Test Scores to Parents (Unstandardized Coefficients)*

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Initial Reading Score</th>
<th>Sex (female)</th>
<th>SES</th>
<th>Hispanic</th>
<th>Asian</th>
<th>Black</th>
<th>White</th>
<th>English at home</th>
<th>Student Age in months</th>
<th>Children with disability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.94 (0.01)***</td>
<td>-0.69 (0.14)***</td>
<td>0.90 (0.13)***</td>
<td>-0.45 (0.24)</td>
<td>0.61 (0.40)</td>
<td>-0.68 (0.34)*</td>
<td>-0.07 (0.34)</td>
<td>-0.64 (0.25)*</td>
<td>-0.03 (0.02)</td>
<td>-1.22 (0.21)***</td>
</tr>
</tbody>
</table>

Note: *p≤.05, **p≤.01, *** p≤.001

Weighted by W4C4P_20
Table 9

Reading Achievement at Spring Kindergarten (Continued)

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Reading instructional time</th>
<th>Phonics</th>
<th>Integrated language arts</th>
<th>Phonics in context</th>
<th>Reading achievement</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of minority student</td>
<td>-0.02 (0.01)**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% of free lunch</td>
<td>0.02 (0.01)**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Private school</td>
<td>1.15 (0.52)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reporting test scores parents</td>
<td>83.92 (21.19)**</td>
<td>0.08 (0.05)</td>
<td>0.07 (0.06)**</td>
<td>0.18 (0.08)*</td>
<td>0.62 (0.51)</td>
</tr>
<tr>
<td>Frequency of ST</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.33 (0.29)</td>
</tr>
<tr>
<td>Average teacher education</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.46 (0.34)</td>
</tr>
<tr>
<td>Average teaching experience</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.06 (0.02)**</td>
</tr>
<tr>
<td>Location of school</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.18 (0.14)</td>
</tr>
<tr>
<td>Reading instructional time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.01 (0.01)*</td>
</tr>
<tr>
<td>Phonics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-1.08 (0.52)*</td>
</tr>
<tr>
<td>Integrated language art</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.59 (0.28)</td>
</tr>
<tr>
<td>Phonics in context</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.61 (0.28)*</td>
</tr>
<tr>
<td>Residual variance</td>
<td>23651.48 (1623.60)</td>
<td>.14 (.04)</td>
<td>26 (.04)</td>
<td>.39 (.04)</td>
<td>-</td>
</tr>
</tbody>
</table>

Note: *p ≤ .05, **p ≤ .01, ***p ≤ .001
Weighted by W4C4P_20
Table 10

*Descriptive Unweighted and Weighted Statistics for Reading achievement*

<table>
<thead>
<tr>
<th>Reading</th>
<th>Unweighted</th>
<th></th>
<th></th>
<th></th>
<th>Weighted</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Min</td>
<td>Max</td>
<td>Mean</td>
<td>SD</td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>K (fall)</td>
<td>37.98</td>
<td>9.85</td>
<td>21.51</td>
<td>90.35</td>
<td>37.73</td>
<td>9.56</td>
<td>21.51</td>
<td>90.35</td>
</tr>
<tr>
<td>K (spring)</td>
<td>50.29</td>
<td>11.72</td>
<td>21.95</td>
<td>90.35</td>
<td>50.06</td>
<td>11.45</td>
<td>21.95</td>
<td>90.35</td>
</tr>
<tr>
<td>Grade 2 (fall)</td>
<td>56.17</td>
<td>13.67</td>
<td>25.17</td>
<td>93.40</td>
<td>56.51</td>
<td>13.30</td>
<td>27.56</td>
<td>93.40</td>
</tr>
<tr>
<td>Grade 1 (Spring)</td>
<td>70.18</td>
<td>13.07</td>
<td>25.27</td>
<td>95.13</td>
<td>69.96</td>
<td>12.96</td>
<td>25.27</td>
<td>95.13</td>
</tr>
</tbody>
</table>

*Note:* The longitudinal sample included 29,568 observations from 7392 children nested in 744 kindergartens. Weighted by W4C4P_20
Table 11

Growth Models for Reading Achievement (IRT) from Kindergarten Entry until the Spring of First Grade

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictors of Intercepts</th>
<th>Fixed effects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intercept, $\gamma_{000}$</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>70.27*** (.18)</td>
<td>70.27*** (.22)</td>
</tr>
<tr>
<td>B</td>
<td>-0.69*** (.15)</td>
<td>-0.69*** (.13)</td>
</tr>
<tr>
<td>C</td>
<td>-0.37 (.24)</td>
<td>-0.39 (.24)</td>
</tr>
<tr>
<td>D</td>
<td>-0.67*** (.26)</td>
<td>-0.70** (.26)</td>
</tr>
<tr>
<td>E</td>
<td>0.48 (.31)</td>
<td>0.48 (.31)</td>
</tr>
<tr>
<td></td>
<td>Other race, $\gamma_{050}$</td>
<td>0.35 (.31)</td>
</tr>
<tr>
<td>F</td>
<td>1.10*** (.12)</td>
<td>1.11*** (.12)</td>
</tr>
<tr>
<td>G</td>
<td>Initial reading, $\gamma_{070}$</td>
<td>0.91***(.01)</td>
</tr>
<tr>
<td>H</td>
<td>Disability, $\gamma_{080}$</td>
<td>-1.33 ***(.18)</td>
</tr>
<tr>
<td>I</td>
<td>Home-language, $\gamma_{090}$</td>
<td>-0.49 (.26)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: *p≤.05, ** p≤.01, *** p≤.001
Weighted by W4C4P_20
Grand Mean-Centered
Table 11

Growth Models for Reading Achievement (IRT) from Kindergarten Entry until the Spring of First Grade (Continued)

<table>
<thead>
<tr>
<th>Model</th>
<th>Fixed effects</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictors of Intercepts</td>
<td></td>
<td>Level 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K Frequent ST, $\gamma_{001}$</td>
<td>.36 (.25)</td>
<td>.41 (.27)</td>
<td>.40 (.27)</td>
<td>.50 (.28)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K Test report, $\gamma_{002}$</td>
<td>.20 (.51)</td>
<td>.20 (.53)</td>
<td>.21 (.53)</td>
<td>.27 (.51)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K reading time, $\gamma_{003}$</td>
<td>.01 (.01)</td>
<td>.01 (0.01)</td>
<td>.01 (0.01)</td>
<td>.01 (0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phonics, $\gamma_{004}$</td>
<td>-.47* (.19)</td>
<td>-.47* (.19)</td>
<td>-.47* (.22)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ILA, $\gamma_{005}$</td>
<td>.15 (.23)</td>
<td>.16 (.23)</td>
<td>.15 (.22)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phocntext, $\gamma_{006}$</td>
<td>.33 (.23)</td>
<td>.33* (.23)</td>
<td>.33 (.21)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School locale, $\gamma_{007}$</td>
<td>.28* (.12)</td>
<td>.28 (.14)</td>
<td>.28* (.14)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Private, $\gamma_{008}$</td>
<td>.22 (.55)</td>
<td>.23 (.55)</td>
<td>.17 (.48)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% of minority, $\gamma_{009}$</td>
<td>-.01 (.01)</td>
<td>-.01 (.01)</td>
<td>-.01 (.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

Weighted by W4C4P_20

All continuous variables are grand-mean centered.
Table 11

Growth Models for Reading Achievement (IRT) from Kindergarten Entry until the Spring of First Grade (Continued)

<table>
<thead>
<tr>
<th>Model</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Predictors of Reading</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Slope</td>
<td>Fixed effects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept, $\gamma_{00}$</td>
<td>1.72 *** (.01)</td>
<td>1.72 *** (.01)</td>
<td>1.72 *** (.01)</td>
<td>1.79*** (.01)</td>
<td>1.79*** (.01)</td>
</tr>
<tr>
<td>Level 2</td>
<td>-.08*** (.01)</td>
<td>-.08*** (.01)</td>
<td>-.08*** (.01)</td>
<td>-.08*** (.01)</td>
<td></td>
</tr>
<tr>
<td>Male, $\gamma_{11}$</td>
<td>-.08*** (.02)</td>
<td>-.08*** (.01)</td>
<td>-.08*** (.01)</td>
<td>-.08*** (.01)</td>
<td></td>
</tr>
<tr>
<td>Black, $\gamma_{12}$</td>
<td>-.05*(.02)</td>
<td>-.06**(.02)</td>
<td>-.06**(.02)</td>
<td>-.06**(.02)</td>
<td></td>
</tr>
<tr>
<td>Asian, $\gamma_{13}$</td>
<td>0.03 (.03)</td>
<td>.03 (.02)</td>
<td>.03 (.02)</td>
<td>.03 (.02)</td>
<td></td>
</tr>
<tr>
<td>SES, $\gamma_{14}$</td>
<td>0.12*** (.01)</td>
<td>.13*** (.01)</td>
<td>.13*** (.01)</td>
<td>.13*** (.01)</td>
<td></td>
</tr>
<tr>
<td>Initial reading, $\gamma_{15}$</td>
<td>-.01***(.01)</td>
<td>-.01***(.01)</td>
<td>-.01***(.01)</td>
<td>-.01***(.01)</td>
<td></td>
</tr>
<tr>
<td>Disability, $\gamma_{16}$</td>
<td>-.01***(.01)</td>
<td>-.01***(.01)</td>
<td>-.01***(.01)</td>
<td>-.01***(.01)</td>
<td></td>
</tr>
</tbody>
</table>

Note: *p≤.05, ** p≤.01, *** p≤.001

Weighted by W4C4P_20

All continuous variables are grand-mean centered
Table 11

**Grand Mean-Centered Growth Models for Reading Achievement (IRT) from Kindergarten Entry until the Spring of First Grade (Continued)**

<table>
<thead>
<tr>
<th>Model Predictors of Reading Slope</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K Frequent ST, $\gamma_{001}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.01 (.02)</td>
</tr>
<tr>
<td>Frequent ST x Male, $\gamma_{111}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.02 (.02)</td>
</tr>
<tr>
<td>Frequent ST x Black, $\gamma_{121}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.01 (.03)</td>
</tr>
<tr>
<td>Frequent ST X Asian, $\gamma_{131}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.06* (.03)</td>
</tr>
<tr>
<td>Frequent ST X SES, $\gamma_{141}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.03* (.01)</td>
</tr>
<tr>
<td>Frequent ST X Initial reading, $\gamma_{151}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.01 (.01)</td>
</tr>
<tr>
<td>Frequent ST X Disability, $\gamma_{161}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.02 (.02)</td>
</tr>
</tbody>
</table>

Note: *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$  
Weighted by W4C4P_20
Table 12

*Random Effects for All Models*

<table>
<thead>
<tr>
<th>Model</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance Estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level-one variance ($\sigma^2$)</td>
<td>35.33</td>
<td>35.33</td>
<td>35.33</td>
<td>34.41</td>
<td>34.37</td>
</tr>
<tr>
<td>Intercept variance ($\tau_{00}$)</td>
<td>18.99</td>
<td>19.00</td>
<td>19.01</td>
<td>19.23</td>
<td>19.28</td>
</tr>
<tr>
<td>Slope variance</td>
<td>2.64</td>
<td>2.62</td>
<td>2.56</td>
<td>2.67</td>
<td>2.62</td>
</tr>
<tr>
<td>Error covariance</td>
<td>0.30</td>
<td>0.29</td>
<td>0.28</td>
<td>0.29</td>
<td>0.28</td>
</tr>
<tr>
<td>Deviance statistic</td>
<td>198686.39</td>
<td>198681.33</td>
<td>198657.11</td>
<td>198074.22</td>
<td>198058.62</td>
</tr>
<tr>
<td>Number of parameters</td>
<td>14</td>
<td>16</td>
<td>22</td>
<td>29</td>
<td>36</td>
</tr>
<tr>
<td>Deviance change</td>
<td>5.06</td>
<td>24.22</td>
<td>582.89</td>
<td>15.6</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX B

FIGURES
Figure 1: A Conceptual Model Relating Standardized Tests to Student Achievement:

School Characteristics
- Sector
- Locale (city, suburb, town rural)
- Percent of non-white students
- Percent of free lunch
- Teachers' average education
- Teachers average years of teaching experiences

Student Characteristics
- Gender
- Age in months
- Race/ethnicity
- Socioeconomic status
- Initial reading score
- Home language
- Disability status

School-level testing policies
- Frequency of standardized tests
- Reporting Test scores to parents

School-level Reading Instruction
- Types of instruction
- Amount of instruction

Student Achievement
- Reading Achievement
Figure 2: Graphical Depiction of the Reading Instruction CFA

RCGNZE = Recognizing alphabet and letters; WRTME = Writing own name; MATCH = Matching letters to sounds; MAINID = Identifying the main idea and parts of a story; ORALID = Communicating complete ideas orally; DOPROJ = Doing an activity or project related to a book or story; PREDIC = Making predictions based on text; TEXTCU = Using context cues for comprehension; RETELL = Retelling stories; CONVOC = Controlled vocabulary; PATTXT = Patterned text, PHOTXT = Phonetic text.
Figure 3: Fixed-effects model depicting relations between K frequency of standardized testing policy and teacher’s reading instruction and k reading achievement. This model fit the data well on most indices: $\chi^2$ (Unweighted N = 12,241, df = 201, MLR scaling correction factor = 1.99) = 1465.71, p < .0001; CFI = .95; TLI = .92; RMSEA = .02; SRMR = .01(within), .11(between). 12,241 students nested in 1,067 schools. Solid lines represent significant relations, whereas dashed lines represent non-significant relations. The double dashed line represent, pathway of mediated effect. Unstandardized estimates are above the standardized estimates, which are in parentheses. K: Kindergarten; *p < .05; **p < .01, ***p < .001. The coefficient in parenthesis denotes the standard error.
Figure 4. Fixed-effects model depicting relations between reporting test scores to parents at school-level and teacher’s reading instruction and kindergarten reading achievement. This model fit the data well on most indices: \( \chi^2 \) (Unweighted N = 12,241, df = 201, MLR scaling correction factor = 1.99) = 1492.50, p < .0001; CFI = .95; TLI=.92; RMSEA = .02; SRMR = .01(within), 0.11(between). 12,241 students nested in 1,067 schools. Solid lines represent significant relations, whereas dashed lines represent non-significant relations. The double dashed line represent, pathway of mediated effect. Unstandardized estimates are above the standardized estimates, which are in parentheses. K: Kindergarten; *p < .05; **p < .01, ***p < .001. The coefficient in parenthesis denotes the standard error.
Figure 5: Examples of Individual Reading Growth over Time
Figure 6 Linear Slopes for Monthly Reading Growth
Figure 7 Interaction between Parental income/Occupation/Education and Frequency of Standardized Tests
Figure 8 Interaction between Asian children and Frequency of Standardized Tests