EXAMINING THE RELATIONSHIP BETWEEN FIDELITY OF IMPLEMENTATION AND STUDENT OUTCOMES WITHIN A SCHOOLWIDE READING MODEL

by

ELIZABETH ANN JANKOWSKI

A DISSERTATION

Presented to the Department of Educational Methodology, Policy, and Leadership
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2015

DISSERTATION APPROVAL PAGE

Student: Elizabeth Ann Jankowski

Title: Examining the Relationship Between Fidelity of Implementation and Student Outcomes Within a Schoolwide Reading Model

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Educational Methodology, Policy, and Leadership by:

Dr. Gina Biancarosa Chairperson
Dr. Charles Martinez Core Member
Dr. Keith Hollenbeck Core Member
Dr. Elizabeth Harn Institutional Representative

and

Scott L. Pratt Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2015.

DISSERTATION ABSTRACT

Elizabeth Ann Jankowski

Doctor of Philosophy

Department of Educational Methodology, Policy, and Leadership

June 2015

Title: Examining the Relationship Between Fidelity of Implementation and Student Outcomes Within a Schoolwide Reading Model

The purpose of this study was to make use of indicators of level of implementation collected during the enactment of Oregon Reading First in order to examine whether variation of implementation of the components of the Schoolwide Reading Program predicted better outcomes for students and schools. In particular, the aim of this study was to determine the extent to which each of three different types of measures of implementation fidelity as well as a combined index of these measures explained school-level variance in student improvement in 34 schools participating in the Reading First program. Hierarchical linear modeling was utilized to predict reading performance and growth on oral reading fluency and overall measures of reading performance. Mixed results, at best, were found when analyzing this association. In both second and third grades, one of three implementation indices and a composite total of all three measures were statistically significant but small predictors of oral reading fluency growth. However, this relationship was offset with the removal of one outlier school. Implementation threshold effects are discussed as a possible cause of nullification. No statistically significant relationships were found between implementation fidelity measures and overall reading outcomes directed at reading comprehension. Although not

a focus of the study, school-level demographic characteristics including special education status and limited English proficiency appeared to explain significant differences between schools despite the use of evidence-based practices and strong support for implementation of these practices.

CURRICULUM VITAE

NAME OF AUTHOR: Elizabeth Ann Jankowski

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
Kansas State University, Manhattan
University of Iowa, Iowa City

DEGREES AWARDED:

Doctor of Philosophy, Educational Methodology, Policy and Leadership, 2015, University of Oregon
Master of Science, Special Education, 2004, Kansas State University
Bachelor of Science, 1987, Elementary and Special Education, University of Iowa

AREAS OF SPECIAL INTEREST:

K-12 Literacy Education
Literacy Leadership
Special Education
Multi-Tiered Systems of Support
Teacher Professional Development

PROFESSIONAL EXPERIENCE:

Research Assistant and Project Manager, University of Oregon, Technical Assistance and Consulting Services, 2012-2015

Research Assistant, University of Oregon, Center on Teaching and Learning 2006-2012

Staff Development Specialist and Special Education Consultant, Iowa/Heartland Area Education Agency, 1997-2006

Adjunct Instructor, Kansas State University, Department of Special Education, 1996-1997, 2000-2004

Elementary and Middle School Special Education Teacher, 1987-1996
Iowa, Kansas, Department of Defense Dependents Schools-Germany

GRANTS, AWARDS, AND HONORS:

Co-Principal Investigator and Project Manager for *Project LIFT*, Contract with Federated States of Micronesia, National and State Departments of Education, University of Oregon, 2012-Present

PUBLICATIONS:

Jankowski, E.A. (2003). Heartland Area Education Agency's problem solving model: An outcomes-driven special education paradigm. *Rural Special Education Quarterly, 22* (4), 29–36.

ACKNOWLEDGMENTS

First, my sincere thank you to my advisor, Dr. Gina Biancarosa, for her thoughtful guidance and ongoing support during the completion of this study. My thanks, also, to Dr. Charles Martinez, Dr. Keith Hollenbeck, and Dr. Beth Harn, members of my dissertation committee, for their time and feedback on the preparation of this dissertation.

Thank you also to Dr. Deni Basaraba for her encouragement and insights into Oregon Reading First activities and to Dr. Patrick Kennedy for his assistance with obtaining the data for this study.

I also wish to express my sincere appreciation to Dr. Stan Paine as well as John Inglish, my former supervisors at the University of Oregon, both of whom were extremely supportive of my studies during my time in the doctoral program. Thank you, also, to Dr. Lauren Lindstrom for her logistical support during the final stages of the completion of this manuscript.

To my husband David, for his extraordinary support and patience, and to my late father, Norbert Jankowski, an educator himself, who modeled the value of life-long learning.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

School literacy reforms have taken many forms and directions over the past 30 years (Correnti & Rowan, 2007; Rowan, Camburn & Barnes, 2004). Notably among these reforms was the passage of the No Child Left Behind Act (NCLB; 2002) and establishment of one of its six mandated programs, Reading First. Conducted from approximately 2002-2008, Reading First distributed over $900 million in federal funds to state and local education agencies for use in low-performing schools with well-constructed plans for improving the quality of reading instruction. Reading First's goal was to "ensure that all children in America learn to read well by the end of third grade" (U.S. Department of Education, 2002). The program sought to integrate the essential components of reading instruction into K-3 reading structures of each State and required that programs and instruction within Reading First schools be based upon scientific research.

National Reading First

The U.S. Department of Education (2002) set federal guidelines and requirements for Reading First plans. The initial application plans for Reading First funding required states to describe the state educational agency's plan for implementing the Reading First program. This plan required states to include specific components using the following language:

- Identification of reading assessments with proven validity and reliability – The SEA must describe how it will assist local educational agencies in identifying screening, diagnostic, and classroom-based instructional reading assessments.

- Identification of scientifically based materials and programs – The SEA must describe how it will assist local educational agencies in identifying instructional materials, programs, strategies and approaches that are based on scientifically based reading research.
- Professional development – The SEA must describe how professional development activities supported with Reading First funds will effectively improve instructional practices for reading and ensure that these activities are based on scientifically based reading research.
- Implementing the essential components of reading instruction – The SEA must describe how funded activities will help teachers and other instructional staff to implement the essential components of reading as identified in the National Reading Panel Report (2000).

Gamse et al. (2008) described the Reading First program as a funding stream that combined national commonalities and local flexibility. The commonalities were reflected in the guidelines to states, districts and schools regarding allowable use of resources, such as those set forth in the application guidelines listed above. The flexibility was that states could make local decisions about the specific choices within given program categories such as which materials, reading programs, assessments, and professional development providers that would be used within their state plans. No mention was made in the guidelines relative to service delivery options for reading programs, including the use of multi-tiered systems of support. Hence, states had flexibility in how the Reading First program would be delivered within each local Reading First school. The activities, programs, and resources that were likely to be

implemented across states and districts would therefore reflect both national priorities as well as local interpretations.

Oregon Reading First

Given local flexibility, it is not surprising that states varied in how Reading First plans were carried out (McKenna & Walpole, 2010; U.S. Department of Education, 2011). The State of Oregon began funding its first cohort of Reading First schools in 2003-2004 and a second cohort of schools two years later, in 2005-2006 (Baker et al., 2011; Sanford, Park, & Baker, 2013). Oregon chose to implement Reading First through a specific framework of early reading instruction called the Schoolwide Reading Model. The Schoolwide Reading Model, as described by Simmons, Kuykendall, King, Cornachione, & Kameenui (2000), can be construed as a multi-tiered system of supports (MTSS) framework with many elements similar to Response to Intervention (RTI), although the overall goal of implementation was not for purposes of identifying students for special education services, but rather schoolwide reading improvement. Multi-tiered systems of supports generally involve collecting valid and reliable assessment data to inform instruction, using differentiated and multi-tiered instruction, and promoting the use of evidence-based practices and programs.

Two in-depth studies analyzed the results of the Oregon Reading First Program. Baker et al. (2011) studied the impact in general of Oregon Reading First on student reading outcomes. Using the hypothesis that outcome strength of large-scale reform is dependent upon the number of years of implementation (Borman, Hewes, & Overman, 2003), these researchers examined the question of whether student outcomes in schools that were in their third year of implementing Oregon Reading First (identified as Cohort

A) were higher than student outcomes in the second cohort of schools that were in their first year of implementation (identified as Cohort B). Results indicated that the cohort of schools with the most experience implementing Reading First were superior in every grade, K-3, on both formative and summative measures of student reading performance. Sanford et al. (2013) also examined the association between the amount of experience schools had with Oregon Reading First and the reading growth of second and third grade students, with an emphasis on students in special education. Hierarchical Linear Modeling (HLM; Raudenbush & Bryk, 2002) was utilized to predict reading performance and growth on oral reading fluency as a function of time of year, disability status, and amount of experience with the Reading First program. Additionally, a multilevel model was used to predict students' performance on the Stanford Achievement Test-10 (SAT-10) at second grade and the Oregon Assessment of Knowledge and Skills (OAKS) at third grade. Students in more experienced Reading First schools made greater gains on oral reading fluency across second and third grades regardless of their special education status and performed better on measures of reading comprehension in third grade when controlling for initial starting point. In sum, results in both of these studies aligned with the researchers' original hypotheses that students in schools participating in Oregon's Reading First program for longer periods of time made greater reading improvement.

Significance of Fidelity of Implementation

While these two previous studies analyzed the overall reading growth of students within the Oregon Reading First program and, thus, the use of the Schoolwide Reading Model as a framework for improving reading achievement, the assumed positive relationship between the fidelity or level of implementation of the Reading First program

and growth patterns for students within these schools has not been explored. Measuring the relation between implementation fidelity and student outcomes is warranted for a number of reasons. Chief among these is (a) to gain confidence that the observed outcomes of Reading First can indeed be attributed to the program, (b) to gain an understanding of how the quality and extent of implementation of various components of Reading First program implementation potentially affected school and student outcomes, and (c) to add to the growing larger research base on the topic of implementation science and multi-tiered systems of supports. Each of these will be briefly discussed.

When garnering confidence for causal inference, in this case the outcomes within Oregon Reading First, for both theoretical and practical reasons, randomized experiments are the most preferred methodology for assessing treatment effects. Random assignment allows effect estimates that are unbiased, that is, where the expectation of the effect equals the effect in the population (Shadish, 2011; Shadish & Ragsdale, 1996). In the case of Oregon Reading First, randomized control studies were not possible, although statistical analyses were conducted to test for equivalence of Reading First cohorts (Baker et al., 2011) and propensity scoring was used to create groups that were matched on a host of covariates related to student identification for special education in another study (Sanford et al., 2013). In the absence of randomized experiments, using implementation data to predict student outcomes strengthens the ability to make causal claims about the effects of a program (Crawford, Carpenter, Wilson, Schmeister, & McDonald, 2012; Dusenbury, Brannigan, Falco, & Hansen, 2003; Fullan & Pomfret, 1977; O'Donnell, 2008). In general, understanding the contribution of implementation fidelity to student outcomes increases confidence in the validity of reported findings and helps support the

claim that the observed findings can be imputed to the intervention or program. As stated by Berman and McLaughlin nearly 40 years ago, "The bridge between a promising idea and its impact on students is implementation; however, innovations are seldom implemented as planned" (p.349). Without examining fidelity of implementation, it is unclear whether Reading First is causally responsible for the positive effects observed, when in fact some other cause or combination of causes could be responsible for observed effects.

A second reason for examining implementation data is to determine how variations in program implementation might have contributed to variations in student and school outcomes. Collection and reporting of fidelity data in research reports is critical for determining why interventions succeed or fail (Dusenbury, et al., 2003). If schools within Reading First were more or less successful depending upon the amount and/or type of intervention among various components of the Schoolwide Reading Model, this is important information. Speaking in terms of MTSS as a whole, Glover (2010) emphasized that successful implementation of RTI requires examination of specific components of service delivery by collecting and responding to fidelity-monitoring data. Additionally, in a review of lessons learned from the larger federal Reading First program in its entirety, Kovaleski and Walpole (2010) suggested that in future evaluations of federal project initiatives, levels of implementation be gauged to determine whether impact varies with respect to fidelity and, if so, what factors have proved conducive to higher levels of implementation.

Additionally, there does appear to be a call in the larger context of education for research on implementation. The relation between implementation fidelity and student

outcomes is widely understudied, and, in particular, the research on implementation fidelity for school-wide systems of supports is just in its beginning stages (Crawford, Carpenter, Wilson, Schmeister, & McDonald, 2012; Harms, 2010; Pas & Bradshaw, 2012). Results of this study will therefore add to the growing literature base in understanding fidelity of implementation as it relates to multi-tiered systems of supports as a whole.

Purpose and Significance of the Study

That being said, the purpose of the current study was to make use of indicators of level of implementation collected during the enactment of Oregon Reading First in order to examine whether variation of implementation of the components of the Schoolwide Reading Program predicted better outcomes for students and schools. In particular, the aim of this study was to determine the extent to which each of three different types of measures of implementation fidelity, as well as a combined index of these measures collected during the implementation of Oregon Reading First, explained school-level variance in student improvement. Before answering these questions, a review of the literature on the Schoolwide Reading Model as well as previous research on fidelity of implementation is presented.

CHAPTER II

LITERATURE REVIEW

Fixen and colleagues (2005) suggested separating the evidence of effectiveness of practices and programs from the *implementation* of evidence-based practices and programs. As they noted, a critical notion to understand is that evidence of the effectiveness of certain practices or programs for specific populations helps us choose *what* to implement. However, evidence of the effectiveness of certain practices and programs does not mean the practices or programs will be implemented successfully, as researchers cannot assume that an intervention was implemented as planned. As a result, they argued that outcomes need to be evaluated within the context of implementation in order to reach causal conclusions. Based upon this prior work, Fixen, Blase, Metz, and Van Dyke (2013), in an article discussing the difficulties of moving evidence-based practices into routine practice, proposed a formula for successful implementation of evidence-based programs as follows:

Effective Interventions × *Effective Implementation* = *Improved Outcomes*

This formula provides the theoretical context for the literature review that follows. The review will initially discuss the evidence base around the major components of the Schoolwide Reading Model as used within Oregon Reading First. This discussion is followed by a section focused on the definition of implementation fidelity as used within the current study and key understandings around implementation science in general. Next, the existing research base on methods of measuring implementation fidelity as it relates to multi-tiered systems of support and student outcomes will be reviewed. Specific methods of measuring implementation fidelity as used within Oregon Reading

First will also be described. The literature review will conclude with a presentation of research questions querying the relationship between fidelity of implementation to the Oregon Reading First Schoolwide Reading Framework by Reading First schools and student outcomes within those same schools.

Essential Content and Evidence of Effectiveness of the Schoolwide Reading Model

Baker et al. (2011) listed seven key essential elements of the Schoolwide Reading Model (SWRM), as originally described by Kame'enui, Simmons, and Coyne (2000), that were used as the framework for the Oregon Reading First program. These elements also form the foundation for most multi-tiered systems of support in reading currently used across many states and local districts within the United States. These elements include:

1. Schoolwide priorities and practices focus on the essential content in beginning reading development: phonological awareness, alphabetic understanding (i.e., phonics), reading fluency, vocabulary, and comprehension.
2. Reliable and valid assessment data are used to inform instructional practices.
3. Protected and sufficient time is allocated to reading instruction to make sure students reach key reading goals and benchmarks.
4. High-quality implementation of research-based instructional programs is emphasized.
5. Differentiated, multi-tiered instruction provides supports based on individual student need.
6. School-level leadership uses student data to support effective classroom instruction and focuses on sustained, effective implementation.

7. High-quality professional development drives ongoing efforts to continuously improve the quality of reading instruction and student achievement. (Baker et al, 2011, p. 311)

Baker et al. reported that Oregon slightly modified these elements to comply with National Reading First specifications. They cited the requirements of (a) a minimum of 90 minutes of daily literacy instruction for all students that was protected from interruptions in the school schedule, and (b) a comprehensive reading measure at the end of each grade to determine if students were reading at grade level. A brief literature review supporting each of the seven elements of the Schoolwide Reading Model follows.

One of the key elements of the SWRM is the use of a curriculum based upon the essential content of beginning reading instruction as identified by the National Reading Panel (NRP) (2000), including phonological awareness, alphabetic understanding, reading fluency, vocabulary and reading comprehension. A starting point for the Panel's recommendations was *Preventing Reading Difficulties in Young Children* (Snow, Burns, & Griffin, 1998), a consensus report issued by the National Research Council and based upon the best judgments of a diverse group of experts in reading research and instruction. More recently, Stahl and McKenna (2006) presented an updated review of the research on literacy learning that provides additional support for and extends the knowledge base of these five key instructional areas. A definition and brief summary of the research for each component of essential content in beginning reading follows.

Essential Components of Reading

Phonological awareness. Phonological awareness refers to the sensitivity to the sound structure of words (Shanahan, 2005). This term has been assessed and also defined

by many different tasks across many years (Adams, 1990; Shanahan, 2005; Stahl & Murray, 1994; Torgesen & Mathes, 2000). Among these tasks, those involving phonemic awareness, a subskill of phonological awareness, have received the most attention. Phonemic awareness refers to the ability to focus on and manipulate individual phonemes, the smallest distinguishable unit of sound related to meaning, in spoken words (NRP, 2000). In addition to the National Reading Panel Report (2000), a number of other studies provide support for the relationship between instruction in phonemic awareness skills and later reading ability (Melby-Lervag, Lyster, & Hulme, 2012; Scarborough, 2001; Shanahan, 2005), and the theory that phonological processing deficits appear to be the fundamental problem of individuals with reading disabilities (Park & Lombardino, 2012; Siegel, 1993; What Works Clearinghouse, 2012; Wilson & Lesaux, 2001). In sum, historical as well as more recent research indicate that instruction in phonological awareness skills can help support development of early reading skills by preparing children to make the link between sounds and letters.

Alphabetic understanding. A second essential element of reading instruction identified by the NRP is alphabetic understanding. Alphabetic understanding consists of two parts: (a) the alphabetic principle that print maps to the sounds of speech, and (b) the understanding of how letter strings can be phonologically recoded into corresponding sounds and blended to form words (Adams, 1990; Ehri, 2005). A substantial body of knowledge has developed over the past 30 years supporting instruction of alphabetic understanding as critical for students to learn in order to read well (e.g., Adams, 1990; Cheatham & Allor, 2012; Ehri, 2003; Good III, Simmons & Smith, 1998; Snow, Burns & Griffin, 1998; Shanahan, 2006; Swanson, 2008).

This research base extends across many types of learners. For example, in a meta-analysis by Jeynes (2008), a significant positive relationship between phonics instruction and the academic achievement of urban minority elementary school children resulted in medium overall effect sizes. Additionally, the National Literacy Panel on Language-Minority Children and Youth concluded that the same principles of systematic and explicit phonologically based interventions that undergird instruction for English-proficient students also appear to benefit English language learners' (ELLs) literacy development (August & Shanahan, 2006). An update to this report (August & Shanahan, 2010), which included an additional 20 experimental and quasi-experimental studies, provided further confirmation of this conclusion. Research also indicates that intensive instruction of the alphabetic principle for an extended duration can significantly improve outcomes for students identified at-risk for reading difficulties and students with disabilities (Hudson, Torgesen, Lane, & Turner, 2012; Vaughn, 2014; Wanzek & Vaughn, 2007). Collectively, a strong research base exists for including alphabetic understanding as an essential component of instruction for students learning to read.

Oral reading fluency. Oral reading fluency is a third essential component of early reading instruction and is defined as the ability to read text aloud with speed, accuracy, and proper expression (NRP, 2000). A review of the literature encompassing studies over several decades indicates that oral reading fluency relates positively and differentially to reading performance (Fuchs, et al., 2001; NICHD, 2000; Pinnell, Pikulski, Wixson, Campbell, Gough, & Beatty, 1995; Therrien, 2004; Wise et al., 2010). Oral reading fluency has shown to be effective in predicting performance on general proficiency reading and comprehension measures including high-stakes assessments

(Hunley, Davies, & Miller, 2013; Reschly, Busch, Betts, Deno, & Long, 2009; Wanzek, Roberts, Linan-Thompson, Vaughn, Woodruff, & Murray, 2010; Yeo, 2010) . Although there is strong support overall for the relationship between oral reading fluency and comprehension, this relationship can be moderated by characteristics of the subjects being assessed. For example, this relationship may have differential effects depending on grade level assessed (Yovanoff, Duesbery, Alonzo, & Tindal, 2005), EL status (Baker, Park, & Baker, 2012; Lesaux & Kieffer, 2010; Quirk & Beem, 2012), and disability status or type of disability (Chard, Ketterlin-Geller, Baker, Doabler, & Apichatabutra, 2009; Wanzek, Al Otaiba, & Petscher, 2014).

Vocabulary. The NRP lists vocabulary as a fourth essential component of reading. In simple terms, vocabulary is the knowledge of meanings of words. In the context of reading, vocabulary serves as the bridge between the word level processes of phonics and cognitive processes of comprehension (Hiebert & Kamil, 2005). The effects of vocabulary instruction as they relate to reading comprehension are positive and have appeared across a number of years (Elleman, Lindo, & Compton, 2009; Freebody & Anderson, 1983; Schmitt, Jiang, & Grabe, 2011; Stahl and Fairbanks, 1986; Verhoeven & Van Leeuwe, 2008). Researchers have determined this relationship starts early with the development of oral language and extends over grade levels (Cunningham & Stanovich, 1997; Hart & Risley, 1995; NICHD Early Child Care Research Network, 2005; Sparks, Patton, & Murdoch, 2014). As noted by Wagner and Miros (2010), it appears a complex system of direct, indirect, reciprocal, and correlational relationships between vocabulary and reading comprehension do exist. Whether the relationship is

direct or indirect, evidence supports the development of students' oral language and reading vocabulary in order to reach the overall goal of reading comprehension.

Reading comprehension. Using the cognitive conceptualization of text comprehension that reading is purposeful and active (Kintsch and van Dijk, 1978) and the idea that comprehension can be improved by teaching students to use specific cognitive strategies or to reason strategically when they encounters barriers to comprehension, the NRP, after a review of the research, determined that the direct teaching of reading comprehension, particularly reading comprehension strategies, benefit children (NRP, 2000). Hence, reading comprehension instruction was the fifth essential element of reading instruction. The National Assessment Governing Board (2006) defined proficient reading comprehension as the ability to demonstrate an overall understanding of the text and to extend the ideas in the text by making inferences, drawing conclusions, and making connections to their own experiences. The idea supporting explicit instruction of text comprehension is that comprehension can be improved by instructing students to use specific cognitive strategies or to reason strategically when they encounter barriers when reading (NPR, 2000).

A number of literature reviews and meta-analyses summarizing instructional research support the idea that instruction in reading comprehension strategies contributes to improved reading comprehension (Block & Duffy, 2008; Duke & Pearson, 2002; Snow, 2002). There is strong support for use of reading comprehension instruction for struggling students and students with disabilities (Berkeley, Scruggs, & Mastropieri, 2010; Edmonds, Vaughn, Hjelm, Reutebuch, Cable, & Tackett, 2009; Gajria, Jitendra, Sood, & Sacks, 2007; Gersten, Fuchs, Williams Baker, 2001). Two caveats of note in the

literature include the findings that the methods used to teach these strategies do make a difference in reading comprehension (Dewitz, Jones, & Leahy, 2009; Duffy et al, 1986; Duke & Pearson, 2002; Kim, Linan-Thompson, & Misquitta, 2012), and strategies can have differential effects on different groups of learners (McMaster, Espin, & van den Broek, 2014; McMaster, van den Broek, Espin, White, Rapp, Kendeou, et al., 2012). McMaster et al. (2012) suggested that identifying subgroups is important in developing and evaluating the effectiveness of reading comprehension interventions. Additionally, differential effects have been found depending upon the reading comprehension strategy being taught (Berkeley, Scruggs, Mastropieri, 2010; Melby-Lervag, & Lervag, 2014).

Reliable and Valid Assessment Data

In addition to the use of a curriculum based upon the essential content of beginning reading instruction, a second key component of the Schoolwide Reading Model framework is the use of valid and reliable assessment practices to inform instruction. Both content validity and consequential validity are particularly important in assessment. Content validity refers to evidence of content relevance, representativeness, and technical quality; consequential validity is refers to the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use (Messick, 1995). Accurate identification of students at-risk of reading difficulties and the collection of ongoing data to inform instruction have been identified to be a major component of effective multi-tiered systems of support (Hughes & Dexter, 2011; Glover and DiPerna; 2007; Margolis, 2012; Shinn, 2008).

Curriculum-based measurement in reading (CBM-R), as used within Oregon Reading First, has been the subject of extensive research over the past three decades

starting with Marston, Mirkin, and Deno (1984), who sought to pilot an alternative method of referral for student assistance using repeated curriculum-based measurements; these authors concluded that measurement of student performance on curriculum tasks is a feasible and efficacious approach to assessment. A strong evidence base continues to develop around the use of CBM-R to identify students who may be at risk for reading failure and for monitoring progress. For example, Fuchs and Fuchs (2003) reviewed more than 200 empirical studies published in peer-reviewed journals providing evidence of CBM's reliability and validity for assessing the development of competence in reading. They concluded that CBM produces accurate, meaningful information about students' academic levels and growth, is sensitive to student improvement, and when teachers use CBM to inform their instructional decisions, students achieve better outcomes. A significant number of other studies have identified the value of CBM-R for providing reliable and valid screening and monitoring information useful for educational planning and low-stakes decision making (e.g., Christ, 2012; McGlinchey & Hixson, 2004; Reschly, 2009; Stecker, Fuchs & Fuchs, 2007; Stecker, Fuchs, Fuchs, 2008; VanDerHeyden, Witt, Naquin, & Noell, 200l; Wayman, Wallace, Wiley, Tichá, & Espin, 2007).

Protected and Sufficient Instructional Time

Protected and sufficient time for reading instruction is the third element of the Schoolwide Reading Model framework. Early on, Reading First guidance from the U. S. Department of Education (2002) called for a protected, uninterrupted block of time for reading instruction of at least 90 minutes per day. Oregon Reading First schools therefore were instructed to provide at least 90 minutes of uninterrupted instructional

time for all students, with additional daily instructional time provided for students performing below grade level benchmarks. A short review of research as it relates to the importance of instructional time and student achievement is provided below.

Foorman and Torgesen (2001) in an article reviewing critical elements of classroom and small group instruction, asserted there are essentially two ways to increase intensity of preventive instruction in elementary schools. Either the total time in classroom instruction can be increased, or instruction can be provided individually or in small groups. It appears that out of this literature grew the notion of the 90-minute reading block, which is widely recommended as a *starting place* for schools that serve a high proportion of poor and minority students. However, the appropriate amount of time allocated to reading instruction in grades K-3 will vary with the needs of the majority of students. For example, schools that serve a high proportion of students at risk for reading difficulties will likely require a longer block of time devoted to reading instruction than schools that have small numbers of students at risk (Foorman & Torgesen, 2001).

In general, research indicates that increased instructional time is associated with increased achievement. In one example, Dobbie and Fryer (2011) examined charter schools in New York City to identify those elements within schools that had the greatest impact on academic outcomes. The analysis included many traditional measures such as teacher credentials and class size. However, they found that those factors had only weak correlations with student achievement. Instead, their research determined that instructional time, measured as the time students were actually engaged in learning, along with high-dosage tutoring, were much stronger predictors of higher achievement. Harn, Linan-Thompson and Roberts (2008) investigated the role of intensifying instructional

time for the most at-risk first grade readers in schools implementing research-based instructional and assessment practices within multitiered instruction support systems. Results indicated that students receiving more intensive intervention in which the instructional time was nearly doubled compared to the less intensive intervention made significantly more progress across a range of early reading skills. Similar studies that have focused on increasing the amount of instructional time as it relates to achievement have produced significant positive results (e.g., Crawford & Torgesen, 2006; Greenwood, 1991; Simmons et al., 2007).

Differentiated, Multi-Tiered Instruction

A cornerstone of multi-tiered reading models is the use of differentiated instruction and evidence-based interventions designed to either prevent or remediate reading difficulties, often delivered through various tiers of instruction. Oregon Reading First schools using the Schoolwide Reading Model were asked to provide tiered instruction with 90 minutes of reading instruction per day to all students in kindergarten through third grade, with a minimum of 30 minutes of daily differentiated small-group, teacher-directed reading instruction delivered either as part of the 90-minute reading block (typically described as Tier 1 instruction) or, for students who were struggling, in addition to the 90-minute reading block (typically described as Tier 2 instruction). More intensive interventions, typically described as part of Tier 3 instruction, were delivered to students at risk through increased instructional time, research-based intervention reading programs, and/or reduced group size.

A review of the research underscores the efficacy of providing increasingly intensive reading interventions for students experiencing reading difficulties (Algozzine,

Wang, White, Cooke, Marr, et al., 2012; Kamps, Abbott, Greenwood, Wills, Veerkamp, et al., 2008; O'Connor, Harty, & Fulmer, 2005). In an article describing the research base and research needs related to Response to Intervention (RTI) frameworks in primary-grade reading, Denton (2012) reported that a substantial body of converging evidence supports the effectiveness of instructional reading interventions provided to students at risk for reading difficulties in the primary grades. In one of the larger meta-analyses on instructional interventions, Gersten et al. (2009) found strong evidence for providing intensive, systematic instruction on foundational reading skills in small groups to students who perform below grade level, typically three to five times per week for 20 to 40 minutes.

While evidence indicates promising results for students needing supplemental instruction typically found with Tier 2 of multi-tier reading frameworks, the effects of intensified instructional interventions typically delivered within Tier 3 is less clear. Studies of intensive reading interventions provided to students with identified reading disabilities have demonstrated that it is possible to intervene successfully with these students (Denton et al., 2013; Swanson, 1999). Wanzek and Vaughn (2010) reviewed the research on intensive reading interventions to inform Tier 3 instruction for students with reading disabilities. Synthesizing the findings from 18 extensive studies of interventions in kindergarten, first, second, and third grades, Wanzek and Vaughn concluded that early intervention, increasing the intensity of instruction with smaller group sizes, and incorporating multi-component instruction hold promise for planning Tier 3 intervention. However, they also stated there are many unanswered questions requiring additional research examining Tier 3 interventions within fully implemented RTI models.

In another study focusing on intensive interventions for struggling readers, Gilbert et al. (2013) examined the growth of first grade students identified as nonresponsive to general education reading instruction, to a supplemental standardized tutoring program focusing on three of the five essential elements of reading instruction identified by the NRP – phonemic awareness, phonics and fluency – in small-group 45-minute tutoring sessions three times per week. Students who were identified as unresponsive to Tier 2 instruction were randomly assigned to either more of the same Tier 2 tutoring or Tier 3 tutoring with the same content but delivered in a one-to-one setting for 30 minutes five days a week. An analysis of outcomes indicated no differences in change scores between these two groups. The authors proposed that the deficits of students who require Tier 3 intervention "may be better addressed by an individualized or problem-solving approach to RTI in which intervention and assessment are specially designed to meet the needs of each individual student, akin to individualized education programs found traditionally in special education" (p. 151).

School-Level Leadership

The sixth of the seven critical elements of the Schoolwide Reading Model calls for school-level leadership that uses student data to support effective classroom instruction and focuses on sustained, effective implementation. The use of data to make instructional decisions is a relatively new but increasingly important part of the role of educational leaders. Educational leaders at all levels are now called upon to effectively analyze, interpret and apply data findings to make informed decisions in many areas in education, ranging from student instruction to teacher evaluation to commitment of resources (Dadnow, Park, & Wohlstetter, 2007; Fox, 2013; Lachat, Williams, & Smith,

2006; Levin, & Dadnow, 2012) . Data-driven decision-making appears to be a hallmark of good instructional leadership (Creighton, 2001; Fullan, Cuttress, & Kilcher, 2005; Mandinach, Honey, & Light, 2006).

Several themes emerged from the literature on data-based leadership within school districts that were supported and stressed within Oregon Reading First schools. First, leadership personnel should establish a culture of data-based decision-making by making the use of data for decision making non-negotiable and modeling this expectation at school and district levels. Relatedly, principals should help assure teachers are incorporating data into their daily decision making routines (Dadnow et al., 2007; Panettieri, 2006). Second, leadership must be able to provide teachers with timely access to student data through integrated technology systems thus allowing teachers and administrators to use and make sense out the data as needed (Dadnow et al., 2007; Kitchens, 2005; Lachat, Williams, & Smith, 2006). Third, leadership personnel must build capacity at the school level for data driven decision making by investing in professional development of data-informed instruction and providing time for teachers to have collaborative discussions around data both across and between grade levels (Dadnow et al., 2007; DuFour & Marzano, 2009; Panettieri, 2006).

High-Quality Professional Development

The final key component of the Schoolwide Reading Model is professional development that drives ongoing efforts to continuously improve the quality of reading instruction and student achievement. Professional development is defined as the set of knowledge- and skill building activities that raise the capacity of teachers and administrators to respond to external demands and to engage in the improvement of

practice and performance (Elmore, 2002). Evidence suggests that there are strong connections between effective professional development, teacher quality, and student achievement (Darling-Hammond, 1999; Kratochwill, Volpiansky, Clements, & Ball, 2007; Whitehurst, 2002).

Ongoing, sustained and high quality professional development is a consistent theme that resonates across the research on RTI (Denton, 2012; Fuchs & Vaughn, 2012; Griffiths, Parson, Burns, VanDerHeyden, & Tilly, 2007). As noted by Fuchs and Vaughn (2012), differentiation of instruction, a hallmark of RTI, is complex and requires extensive knowledge of reading assessment and instruction. These authors asserted that providing instructional differentiation at the classroom level is often beyond the skill set of even the most proficient teachers, so effective professional development in this area is critical to effective multi-tiered instructional systems. Similar assertions were postulated by Spear-Swerling and Cheesman (2012), who conducted a study of elementary teachers' knowledge regarding Response to Intervention. They reported that, while most participants were familiar with basic features of RTI such as the three-tiered model, they were unfamiliar with the research-based instructional approaches and interventions named in the study.

It is clear that professional development has a strong influence on teacher effectiveness if certain features are in place. Research indicates that short-term or one-session workshops, trainings, conference sessions, etc., have little impact on teacher behavior. Professional development is more effective in changing teachers' practice when it is of a longer duration, involves working with others including peers, and includes job-embedded training – all characteristics of *reform model* professional

development (Darling-Hammond & Richardson, 2009; Garet et al., 2001; Kelleher, 2003). Literacy coaching is also mentioned throughout the literature as one type of professional development that can potentially help teachers bridge the gap between more formal professional development and actual classroom implementation (Carlisle & Berebitsky, 2011; Coburn & Woulfin, 2012; Porche, Pallante, & Snow, 2012).

There is also broad support in the professional development literature for collective participation of groups of teachers from the same school, department, or grade level as opposed to participation of individual teachers from many schools (Darling-Hammond et al., 2009). Among benefits for teachers working collaboratively are increased opportunities to talk over learned knowledge, concepts, and skills as well as collaboratively confronting problems that arise during implementation within their unique context. Establishing professional learning communities (PLC) is one promising approach that can be used for collective professional development (DuFour, 2007; Porche, Pallante, & Snow, 2012). Although studies on the effectiveness of PLCs are somewhat sparse, some evidence of effectiveness is beginning to emerge. Vescio, Ross and Adams (2008) reviewed 11 empirical studies exploring the impact of PLCs on teaching practice and student learning teaching. The collective results of these studies suggested that well-developed PLCs have positive impact on both teaching practice and student achievement. Additionally, Lomos, Hofman and Bosker (2011), in a meta-analysis of studies investigating the effects of professional learning communities on student achievement, reported a small but significant summary effect ($d = .25$, $p < .05$), indicating a professional learning community has the potential to enhance student achievement.

Fidelity of Implementation

Implementation is not simply an extension of planning and adoption processes. It is a phenomenon in its own right (Fullan & Pomfret, 1977, p.336).

As the research review of each the seven key components of the Schoolwide Reading Model as used within Oregon Reading First has just indicated, substantial evidence supports each of these elements. However, effective intervention is only half of Fixen's (2013) formula for improved student outcomes. The other half of the formula calls for *effective implementation* of these elements by practitioners, (i.e., *Effective Interventions* × *Effective Implementation* = *Improved Outcomes*). As Fixsen, Blase, Horner, and Sugai (2009) suggest, "choosing an evidence-based practice is one thing, implementation of that practice is another thing altogether" (p. 5).

Understanding Fidelity of Implementation

Defining fidelity of implementation. A review of the literature indicates there is little consensus across fields regarding the definition of implementation fidelity (e.g., Hagermoser Sanetti & Kratochwill, 2009; Harn, Parisi, & Stoolmiller, 2013; Keller-Margulis, 2012). O'Donnell (2009) suggested that the term *fidelity of implementation* is defined in various ways depending on the type of study (e.g., efficacy or effectiveness research, action research, or program evaluation) and the field of study (e.g., education, mental health or public health). In addition, many terms related to fidelity of implementation are used interchangeably (e.g., treatment integrity, fidelity, intervention integrity, or implementation fidelity) (Keller-Margulis, 2012; Nelson, Cordray, Hulleman, Darrow, & Sommer, 2012; Schulte, Easton, & Parker, 2009). Nuances aside, implementation fidelity is generally defined as the extent to which a program,

intervention, or strategy is used in the manner in which it is designed or intended (Berman & McLaughlin, 1976; De Fazio, Fain, & Duchaine, 2011; O'Donnell, 2008; Sutherland, McLeod, Conroy & Cox, 2013). O'Donnell (2008) asserted that, overall, fidelity of implementation seems to be synonymous with adherence and integrity. This is the definition that will be used for purposes of the current study. That is, fidelity of implementation refers to the extent to which schools adhered to programmic requirements and activities within the seven major components of the Schoolwide Reading Model as stipulated by Oregon Reading First leadership personnel.

Importance of fidelity of implementation. Measuring fidelity to a practice or program is crucial for several reasons. First and foremost, and as suggested earlier, the potential benefit of evidence-based practices is bound by the quality, reach, and maintenance of implementation (Cook & Odom, 2013). Fidelity is paramount to the understanding of any intervention study, as failure to establish fidelity can severely limit the conclusions that can be drawn from any outcome evaluation (Dumas, Lynch, Laughlin, Smith, & Prinz, 2001). Without measuring fidelity of implementation, it is possible to conclude erroneously that observed findings can be attributed to conceptual or methodological underpinnings of a particular intervention rather than success or failure of implementation (Dobson and Cook, 1980; O'Donnell, 2008; Sanettei, & Kratochwill, 2009).

Relatedly, substantial literature points to the importance of fidelity of implementation for the purpose of establishing both external validity and internal validity. Both of these aspects of validity relate to efforts of "scaling up" interventions as discussed below. External validity refers to the extent to which the results of a study can

be generalized to and across populations, settings, and times (Christensen, 2000). External validity is influenced by fidelity because standardized implementation procedures are needed to ensure that an intervention can be replicated in other settings (Allen, Linnan, and Emmons, 2012; Mowbray, Holter, Teague, & Bybee, 2003; O'Donnell, 2008). Internal validity refers to the extent to which outcomes can be attributed to the experimental factors (that is, the essential elements of the intervention) rather than some extraneous or confounding factors (Prohaska & Etkin, 2010). As such, internal validity is threatened when plausible rival explanations cannot be eliminated. Without methodological consideration of the level of fidelity during implementation, researchers may have insufficient evidence to support the internal validity of an efficacy or effectiveness study (Dumas, Lynch, Laughlin, Smith, & Prinz, 2001; Keller-Margulis, 2012).

Prohaska and Etkin (2010) noted that evidence-based programs that address the issues of validity are more likely to be widely used. This is important to the concept of *scaling up* in education. Although defined somewhat differently across disciplines, the term *scaling up* has generally been used to describe efforts to increase the implementation and impact of evidence-based innovations tested in pilot or experimental studies; in turn, this benefits more people and fosters policy and program development on a lasting basis. Researchers point specifically to the need for more empirical research on the association between implementation quality and outcomes when interventions are brought to scale (Greenwood, 2009; Pas & Bradshaw, 2012). Greenwood (2009) states that treatment integrity is key to successful wide-scale application of specific evidence-

based practices and crucial to improving implementation across disciplines more generally.

Finally, looking at fidelity of implementation in another light, information from implementation fidelity measures can help to explain *why* innovations succeed or fail. For example, fidelity of implementation can reveal important information about the feasibility of an intervention, such as how likely it is that the intervention can be carried out as planned (Dusenbury, et al., 2003). Likewise, the process of measuring and analyzing implementation fidelity can also help determine which specific scomponents of the intervention are the most difficult to implement; this, in turn, may explain lack of success with the intervention. Analysis of fidelity data also enables discovery of those elements that make a difference to outcomes, are essential for its success, and whether some elements have a lesser relationship to outcomes. As an example, Benner, Nelson, Stage, and Ralston (2011) sought to examine the extent to which specific elements of fidelity of implementation (i.e., adherence and quality of delivery) enhanced or constrained the effects of a reading intervention for middle school students experiencing reading difficulties. Overall, fidelity of implementation was statistically significant in this study and accounted for 22% and 18% of the variance in the gains in basic reading skills and passage comprehension of middle school students with reading difficulties. Further, the researchers were able to determine through fidelity measurements that two actions in particular by teachers contributed to gains in student achievement above and beyond the contribution of other teacher actions. These actions were effective use of error correction procedures and reteaching when students had not mastered content. In

discussing implications of their results, the authors suggested these two actions become the "look fors" during coaching and administrative classroom visits.

Measuring Fidelity of Implementation

Dimensions of implementation integrity. To adequately measure treatment integrity within research and practice, it is essential to have a conceptual model that defines treatment integrity as a construct (Sanetti & Kratochwill, 2009). As was the difficulty of finding an overall common definition of implementation integrity, the multifaceted nature of fidelity, together with the absence of a unified approach to fidelity within and across research disciplines, have left researchers with little shared basis for measuring and discussing overall dimensions of fidelity of implementation (Century, Rudnick, & Freeman, 2010; Zvoch, 2012). With little shared basis for measuring and discussing, fidelity of implementation researchers have had difficulty comparing findings across studies of particular interventions or accumulating knowledge on fidelity of implementation itself.

One seminal article on implementation fidelity by Dane and Schneider (1998) that is frequently referenced in the field of health and more recently referenced in the field of education (e.g., Benner et al., 2011; Century et al., 2010; Harn et al., 2013; O'Donnell, 2008) describes five aspects or dimensions of measuring fidelity of implementation. Many efforts to develop implementation fidelity measures build upon these following five criteria: (a) adherence – the extent to which specified program components were delivered as prescribed; (b) exposure – an index that may include the number of sessions implemented, the length of each session or the frequency with which program techniques were implemented; (c) quality of delivery – a measure of qualitative aspects of program

delivery that are not directly related to the implementation of prescribed content; (d) participant responsiveness – a measure of participant response to program sessions, and may include indicators such as levels of participation and enthusiasm; and (e) program differentiation – a manipulation check that is performed to safeguard against the diffusion of treatments and to ensure that the subjects in each experimental condition received only planned interventions. (Gerstner & Finney, 2013; Nelson et al., 2012; Sutherland, McLeod, Conroy, & Cox, 2013).

O'Donnell (2008) claimed these five criteria could be considered as divided into two groups including (a) fidelity to *structure* (i.e., adherence, duration), and (b) fidelity to *process* (quality of delivery, differentiation), with participant responsiveness taking on the characteristics of both (cf. Mowbray et al., 2003). Structure simply encompasses the framework for service delivery, and process comprises the way in which services are delivered. Recent literature in the field of education indicates these two broad multidimensional dimensions of fidelity are becoming more common (Crawford et al., 2012; Harn et al., 2013; Odom et al., 2010; Schulte et al., 2009). Harn et al. (2013) described the structural dimensions of fidelity measurement as determining whether important pieces of the intervention established a priori were delivered, citing such examples as adherence to central components, time allocations and intervention completion. Process dimensions, on the other hand, examine the quality of intervention delivery and/or the nature and quality of teacher-student interactions during intervention. In other words, instead of determining whether a component of the intervention simply occurred or not, process fidelity calls for determining to what extent or how well the component was delivered.

Many other models for conceptualizing dimensions of fidelity exist as depicted in Figure 1. Despite different labels for and organization of dimensions across these models, there is overlap among them. For instance, Brandon, Taum, Young, Pottenger, & Speitel (2008) reported that adherence, exposure, and quality are probably the most widely studied aspects of implementation. In their model, adherence has to do with the extent to which the steps and procedures of a program are delivered as intended, exposure has to do with the frequency of program units, and quality is described as the how well a program implements the techniques or methods of the program. Another approach to defining fidelity within the education field uses two organizational categories: (a) structural critical components, and (b) instructional critical components (Century et al., 2010). In an article describing the development of fidelity of implementation measures as its primary goal, Century et al. developed a suite of data collection tools designed to be used across a wide variety of instructional programs. Structural critical components reflect the intervention developers' intentions about the design and organization of the intervention. Instructional critical components, however, reflect the developers' intentions about the participants' (teachers and students) behaviors and interactions as they enact the intervention. Each of the two main categories also has subcategories that further categorize these critical components.

As one final example of differing conceptual definitions of the dimensions of fidelity of implementation, Zvoch (2012) concluded that treatment fidelity has developed as a multidimensional construct that reflects not only the degree to which providers

Figure 1. Examples of Conceptualization of Fidelity of Implementation

Dane & Schneider (1998)		Adherence	Exposure	Participant Differentiation	Quality of Delivery	Participant Responsiveness
O'Donnell (2008; cf. Mowbray et al., 2003)		Structure		Process		
Lynch and O'Donnell (2005)		Structure			Process	Self-perceived Effects of Participants
Brandon, Taum, Young, Pottenger, & Speitel (2008)		Adherence	Exposure		Quality	
Century, J. Rudnick, M., & Freeman, C. (2010)		Structural Critical Components (not necessarily aligned with dimensions of Dane & Schneider)		Instructional Critical Components (not necessarily aligned with dimensions of Dane & Schneider)		
Zvoch (2012)		Adherence			Delivery	Receipt

deliver an intended treatment, program, or service, but also the extent to which targets receive and interact with treatment components. Zvoch stated the delivery, receipt, and adherence/enactment conceptualization serves to outline the broad contours of the fidelity construct and highlight the unique role of providers and recipients in the implementation and use of intervention components. Associated sub-dimensions, including the extent to which a provider delivers the range of treatment components (integrity/adherence) along with the strength (dosage/exposure) and skill of delivery (quality), further identify and characterize the provider role. In this conceptualization, variation in treatment receipt and protocol enactment also matter, as Zvoch asserted an intervention can be delivered with a high degree of skill and integrity, but program participants still may not receive or interact with the treatment as intended. Breakdowns in receipt may occur, for example, when program participants are not engaged during treatment delivery, fail to comprehend

or follow through on treatment-related protocols, and/or intermittently attend treatment sessions.

In summary, a wide variety of terms are used to describe the various dimensions or components of implementation fidelity. Much of the current work emanates from the conceptual work conceived by Dane and Schneider (1998) and those whose seminal thinking came earlier (e.g., Berman & McLaughlin, 1976; Fullan & Pomfret, 1977). As reflected in the discussion above, although the terminology may differ, most models of measuring fidelity call for identifying the critical structural components of an intervention, determining if components were delivered with fidelity (e.g., adherence or structural fidelity), and specifying the degree or quality with which the components were delivered (e.g., integrity or process fidelity).

Methods for collecting treatment fidelity data. Methods used to collect data on implementation fidelity vary and appear to be program or intervention specific. A variety of direct and indirect methods are typically used to determine the level of fidelity when implementing a program or intervention (Gresham, MacMillan, Beebe-Frankenberger, & Bacian, 2000). Direct observation is one of the three more commonly-used techniques for determining level of fidelity, along with self-reports and permanent products (i.e., artifacts) (DeFazio et al., 1977; Durlak & DuPre, 2008; O'Donnell, 2008). Direct observation integrity data is collected by having an observer who is trained in the academic or behavioral intervention observe the teacher and collect real-time data about the accuracy with which the teacher (or other implementer) performs each step. Videotaping and subsequent coding by trained observers can also be utilized. Direct observation requires the various components of treatment be clearly specified in

operational terms so that the occurrence and nonoccurrence of each treatment component can be adequately assessed.

Self-reports, rating scales, interviews and permanent products can also be used to assess the fidelity of implementation and fall within the category of indirect assessment. Self-reports typically consist of asking teachers to use surveys or rating scales to evaluate their own performance (Gresham, et al., 2000). Examples of self-reports include teachers using a 5-point Likert scale with ratings ranging from strongly disagree to strongly agree in order to evaluate the extent to which they implemented each step of an evidence-based strategy, or a school staff rating themselves on the extent they use data to make instructional decisions. Self-reports are often used in measuring fidelity because they are easy to implement, do not require another individual to record data, and are oftentimes cost effective. However, some researchers contend that self-reports may result in inflated accounts of treatment integrity (Century et al., 2010; Gresham et al., 2000). In an example of potential bias in self-reports, Ennett et al. (2011) examined fidelity of implementation issues in a study of middle school teachers implementing a preventative substance use program and conjectured from the results that observational data were less subject to social desirability bias and therefore provided more valid estimates of fidelity than the self-reported data reported in the study. Other studies have noted teacher self-reports on adherence negatively correlating with those of independent observers (Bickman, Riemer, Brown, Jones, & Flay, B., 2009; Dusenbury et al., 2003). As a result of these issues, it is generally not recommended that self-reports be used as the sole means of assessing treatment integrity.

Another frequently used method of measuring fidelity of implementation is the use of permanent products. This method consists of reviewing artifacts from implementation of the intervention to evaluate the fidelity of implementation. For example, permanent products (artifacts) treatment integrity can be used with academic interventions that leave a record of each step of instruction. In this method of permanent product assessment, each step references a permanent product. For instance, if a literacy intervention program step is to spell the words containing a spelling pattern just taught, then the corresponding product would be the presence of those words on program materials. One of the benefits of permanent product data collection is that it does not require significant additional work for teachers or a second person to observe, collect, or review data. Gresham et al. (2000) report permanent product assessment of treatment integrity as being less time consuming, more efficient, less reactive, less likely to be influenced by social desirability, and potentially more accurate than other integrity assessment methods.

All things considered, the methods used to collect data on implementation fidelity vary but typically will involve direct observation of the delivery of an intervention, implementer self-reports, permanent products, interviews and checklists. As noted by Century et al. (2010), not all dimensions of fidelity and not all types of measures are relevant to all interventions. Decisions about what measures of fidelity to use depends on a number of factors including the intervention target (e.g. academic, social, behavioral), the intervention or program recipients (e.g. students, teachers, schools, administrators), the type of data desired (e.g. using a dichotomous score of one or zero or based upon a

range of implementation levels such as a Likert scale) as well as the purpose for which treatment integrity is being collected.

Relationship Between Implementation Fidelity and Student Outcomes

Ostensibly, it would seem correct to assume that in order for an intervention or program to be effective, it must be implemented with the exact specifications designed by its developers. In reality, a review of the literature indicates this is not necessarily the case. As was the difficulty of determining a precise definition for and identifying the dimensions of implementation integrity, the relationship between treatment outcomes and treatment fidelity is filled with subtleties and not so straightforward. A number of factors can impinge on this relationship; each will be briefly discussed below.

General findings. In general, research suggests that higher levels of treatment integrity do result in better outcomes. As evidence, Durlak and DuPre (2008) conducted a systematic and comprehensive review of the literature, examining the impact of implementation on program outcomes. Included in the review were nearly 500 studies summarized in five meta-analyses and an additional 60 studies assessing the impact of implementation on outcomes. Examining effect sizes from individual studies as well as meta-analyses results, Durlak and DuPre concluded that the level of implementation achieved is an important determinant of overall program outcomes and can lead to much stronger benefits for participants. These researchers concluded that achieving good implementation not only increases the chances of program success in statistical terms, but also can lead to much stronger benefits for participants.

Relatedly, Odom et al. (2010) examined fidelity of implementation and its association with different outcome variables using extant data from a large-scale research

study whose main purpose was evaluating an integrated curriculum model designed to promote school readiness. Data was collected from 51 preschool classes located at nationally-dispersed sites. As a whole, data indicated statistically significant positive associations between measures of implementation and several of the child outcome variables. Of particular interest was an interaction showing that children who were lower performing at pretest on literacy measures benefited significantly more from higher levels of implementation than children from the remainder of the group. This occurred even after controlling for race/ethnicity, disability, and status as an English language learner.

The relationship between level of implementation and student outcome measures was also examined by Pas and Bradshaw (2012) using data from a statewide evaluation of Schoolwide Positive Behavioral Interventions and Supports (SW-PBIS). SW-PBIS fidelity, as measured by one of three SW-PBIS fidelity measures, was found to be statistically significantly related with math achievement, reading achievement and truancy such that higher implementation of SW-PBIS was associated with subsequent higher achievement and lower rates of truancy. Similarly, Woodridge et al. (2014) conducted an analysis of extant data from a school-home intervention program with a solid evidence base for achieving positive outcomes with behaviorally at-risk students in the primary grades. The study involved 8,200 students within ten schools in Grades K-8 across the United States. Using HLM regressions, statistically significant effects were found for classroom fidelity, classroom dosage and home intervention dosage. A one standard deviation increase in classroom fidelity increased the intervention effects on a behavior rating scale by 0.29 (p = .01) and a social skills rating scale by 0.25 (p = 0.04). Additionally, a one standard deviation increase in classroom dosage (intervention days)

increased the intervention effect on academically-engaged time by 0.54 ($p < .01$). As to the home component of the intervention, one standard deviation increase in dosage (at home by parents) increased the intervention effect of academically-engaged time at school by 0.36 ($p = 01.$)

Factors influencing the relationship. A deeper review of the literature indicates, however, that the relationship between treatment fidelity to outcomes can vary due to a number of intervening factors including context of implementation, dimension of fidelity assessed, validity and reliability of the fidelity measure, and time-points of fidelity measurement. Additionally, much has been written about the cut-off point for determining what amount should be considered as an adequate amount of fidelity. There is some disagreement among researchers about how much adaptation is allowed without compromising the intervention (Harn et al., 2013; Nelson et al., 2012; Ogden & Fixen, 2014). These issues will be addressed below.

Contextual factors. Although not the focus of this study, an entire body of work studying methods to promote the systematic implementation of research findings and other evidence-based practices has developed and falls under the concept of implementation science (Ogden & Fixen, 2014). Researchers within this field suggest that particular contextual factors may influence quality of implementation as well as eventual outcomes of an intervention. One example is Mihalic and Irwin's 2003 systematic analysis of the factors that could potentially help or hinder implementation efforts of eight violence prevention programs within 42 dissemination sites across the country.

Mihalic and Irwin were interested in studying the influence of human- and systems-level factors that challenge the successful implementation of programs. To make this determination, the researchers developed a set of implementation factor scales designed to include many of the contextual features that seemed to affect implementation success across the program, community, staff, leadership, and agency within which the program was being implemented. For example, within their Ideal Agency Characteristics Scale, the scale included such items as administrative support, political climate and communication, while the Ideal Program Characteristics Scale included such items as staff buy-in, motivation and the hiring pool of available staff. At the end of the second year of implementation, site administrators were asked to rank each item on a Likert-type scale relative to its role in affecting program success. Four dependent variables were used within the study to study this relationship including adherence to core components of the program, percentage of core program components achieved, dosage, and sustainability. Findings from a regression analysis indicated that quality of technical assistance, ideal program characteristics (selecting programs that match the local needs and that are consistent with the stated goals or mission of the school, agency, or community), limited staff turnover, and support from the local community were among the most important facilitators of strong implementation. In sum, data indicated these particular contextual factors can influence quality of implementation and, potentially, its relationship with overall student outcomes.

Referring back to Durlak and DuPre's (2008) previously-cited review of the impact of implementation on outcomes, a second purpose of their study was to determine if and what contextual factors may affect quality of implementation. Durlak and DuPre

analyzed quantitative and qualitative data from over 80 studies on factors affecting the implementation process. A factor was determined to be significant only if it was related to implementation in at least 5 of the 80 research articles and if findings were consistent in the more rigorously conducted investigations. For quantitative studies, this generally meant the use of multiple as opposed to single case studies, prospective rather than retrospective designs, and multiple versus single methods of data collection.

Data from the study identified 23 contextual factors that influenced the level of implementation and, potentially, the outcomes obtained within these studies. These contextual variables fell within five categories of the researchers' ecological framework for successful implementation and included: (a) *community level factors* such as funding, politics, current educational theory and research; (b) *provider characteristic*s, e.g., the perceived need for the program, general skill proficiency, and sense of self-efficacy; (c) *characteristics of the innovation*, e.g., its complexity and its compatibility with the host institution and staff; (d) *organizational capacity* such as work climate, leadership, shared vision and the ways that decisions, communication and problem-solving are enacted within the organization; and (e) *factors related to prevention support systems*, e.g., initial pre-program training and ongoing support and consultation after the program is launched. Durlak (2010) summarized these findings by stating that, because implementation is important to outcomes, it is critical to understand the conditions for achieving effective implementation.

Relatedly, Harn et al. (2013) discussed the need to balance fidelity with contextual fit, citing a number of contextual factors that can moderate fidelity level including teachers' general instructional philosophies, instructional leadership, and

teacher experience. Importantly, the authors asserted that, because every educational environment is unique based upon its own context, matching interventions to the features of that context is key to ensuring a program is successfully implemented and sustained. As a result, researchers need to develop programs that can be adapted to match ever-changing school contexts and student populations.

Dimensions of fidelity. There is also evidence that suggests the method, type and/or dimension of fidelity utilized to document implementation can influence the approximated relationship between fidelity and outcomes (Century et al., 2010; Hirschstein, Edstrom, Frey, Snell, & MacKenzie, 2007; Pas & Bradshaw, 2012; VanDerHeyden, 2012). Indications are that most often, measures take more of a compliance or adherence approach to fidelity. Drake et al. (2001) report that, most typically, scales are developed to quantify fidelity or measure the adherence of an intervention with the model on which it is based through, for example, a checklist. While discussing the differences between measuring structural versus process measurements of fidelity, Mowbray et al. (2003) suggest this may occur because, among other reasons, process criteria often necessitates more time and effort, be more costly and more likely to be less reliable. Mowbray et al. also state, however, that while process criteria may be more difficult to measure reliably, they may be significant as far as program effects. Therefore, fidelity criteria should include aspects of both structure and process.

As an example of the possible influence of the type of dimension measured on program outcomes, Pas and Bradshaw (2012), in a previously-cited study measuring the association between implementation and outcomes in a statewide scale-up of schoolwide positive behavior intervention and supports (SW-PBIS), used three different measures to

assess the level of implementation of SW-PBIS. It was hypothesized that higher levels of implementation would be associated with higher levels of achievement and lower rates of truancy and suspensions. Interestingly, only one of the three fidelity measures were significantly related to the outcome measures. The authors hypothesized that the differences detected in predictive validity may have been the result of the three measures assessing slightly different aspects of SW-PBIS implementation. They also noted that the fidelity measure which did have statistically significant associations with several outcome measures was the only measure completed by an outsider to the schools. The mixed findings within the study between type of fidelity measure and student outcomes demonstrated how the choice of an implementation measure and dimension measured influenced the pattern of findings.

As another example, Crawford et al. (2012) examined the relationship between fidelity of implementation and student outcomes in a computer-based middle school mathematics curriculum involving nearly 500 students. Fidelity to implementation was measured via two broad constructs: fidelity to structure and fidelity to process. The authors categorized three measures as fidelity to structure and included: (a) total time in intervention, (b) concentration of time in the intervention, and (c) teacher adherence to and student engagement with the program as measured through direct observation. Fidelity to process was measured through use of a rating scale containing process variables essential in the delivery of computer-based instruction such as teacher communication and classroom management. Using a two-level HLM model of analysis, results showed that increased fidelity to structure related significantly to higher outcomes in student posttests, whereas fidelity to process demonstrated no significant increase in

outcome measures. Taken altogether, these findings again illustrate how the choice of an instrument and the dimension of fidelity that it measures can influence its relationship with intervention or program outcomes.

Validity and reliability of fidelity measures. An analysis of the literature on the factors that may influence the relationship between implementation fidelity and outcomes also points to the importance of examining reliability and validity of the measures used to obtain implementation fidelity scores (O'Donnell, 2008; McKenna, Rosenfield, & Gravois, 2009; McLeod, Southam-Gerow, and Weisz, 2009; Nelson, et al., 2012; Sheridan Swanger-Gagne, Welch, Kwan, & Garbacz, 2009). Unfortunately, this also appears to be an area that is currently understudied (Fixen et al., 2005; Mowbray et al., 2003). As Ogden and Fixen (2014) recently noted, there is a great need for the development of instruments which operationalize and standardize the measurement and analyses of implementation processes and outcomes.

Quality of implementation (process) is more subjective than adherence and dosage (structure) and, therefore, more difficult to define and measure (Fagan, Hanson, Hawkins, & Arthur, 2008). Quality is generally recorded as a continuous variable, and most often examined through observations. To assist with reliability, observations require the comprehensive training of observers and evidence of acceptable inter-observer reliability (Brandon, Lawton, & Harrison, 2014). Zvoch (2009) discussed the significance of reliability of fidelity indices by asserting that lack of systematic and extensive data on the reliability of classroom observations that do not include more than one observer could be considered as a limitation of an evaluation of an implementation.

In other words, the quality of inter-rater reliability should be considered when examining the relationship between direct observations of fidelity and student outcomes.

Nelson et al. (2012) contended that because of the uncertainty of efficacy of many measures of implementation fidelity, reliability can be enhanced by using multiple methods for measuring implementation of the individual components of an intervention. They note that a self-report may be less reliable and more biased method of reporting teacher practices, but it may also allow the researchers to get at elements of implementation that classroom observations may not be able to detect reliably. In such a case, using multiple measures or a single measure with multiple items with minimally correlated measurement error better allows for measurement of the underlying construct with more reliability.

Content validity is also a concern discussed in the literature with regard to implementation fidelity measures. Simply described, content validity refers to how accurately an assessment or measurement represents the various aspects of the specific construct in question. For a fidelity measure to produce data useful in making decisions relative to the presence or absence of evidence-based practices, the measure must be sensitive to the key observable dimensions of the intervention (Greenwood, 2009; O'Donnell, 2008). In order to gain the most information from measures of implementation fidelity, the measures must show if the component structures were carried out *and* the degree to which these components were delivered. In other words, researchers will know that more or less of the intervention is present, and which aspects are missing and need improvement.

Time points of measurement. Finally, the relationship between fidelity of implementation and outcomes can vary according to the time point at which fidelity is measured, although there does appear to be mixed results regarding the influence of this variable. Harn et al. (2013), postulated that when schools initially implement an intervention, fidelity of implementation may be uncharacteristically low due to interventionists attempting to understand how the program works in general, how the program components interact with their particular students, and other novice implementation considerations. This was in fact the finding of Harms (2010) in a study designed to examine the process of implementing an integrated three-tier model and explore the relation between implementation fidelity and student outcomes. Results showed that average implementation fidelity scores improved over time, with the most amount of implementation growth occurring during the first year of implementation. Similar findings were made by Sanetti and Kratochwill (2011). In their review of school-based consultation studies, the authors found strong evidence that a vast majority of teachers implement interventions with low levels of treatment integrity within 0–10 days of initiating an intervention and then gradually improve over time.

In contrast to these findings, however, Noell et al. (2005) examined teachers' implementation of treatment plans following expert consultation. Within this study, interventions were implemented with 45 elementary school students referred for consultation and intervention due to academic concerns, challenging behaviors, or a combination of the two. Teachers of these students were assigned to one of three weekly consultation conditions to discuss the student interventions over a three-week period. Surprisingly, intervention treatment integrity was somewhat higher the first week of

implementation compared to the second and third week. Treatment integrity the first week was statistically significantly higher from integrity for Week 2 ($t = 3.4$, $df = 44$, $p = .001$) and Week 3 ($t = 3.7$, $df = 44$, $p = .001$).

In a related inquiry, Zvoch (2009) sought to examine the extent that program adherence varied initially and over time within and between 52 Head Start classrooms implementing two early childhood literacy curricula. In this study, research staff collected implementation fidelity data at three points in time across the year. Using a multi-level growth curve analysis, results indicated that fidelity to a program protocol did significantly vary over time. Zvoch suggested that a snapshot of adherence at one point in time (even if a reliable and valid observation) may not be a good indicator of past or future adherence levels and contended that evaluators would be well served by the repeated collection of implementation data.

Acceptable Levels of Treatment Integrity

One of the more complex issues related to the concept and measurement of treatment integrity revolves around an acceptable level of treatment integrity, including how much adaptation is allowed without compromising an intervention. There appears to be an inherent tension that often exists between researchers and practitioners, with researchers wanting practitioners to implement the curriculum exactly as it was designed, and practitioners wanting to modify components of the practice to fit their context (Odom, 2009). A review of the research indicates that, while no one states 100% adherence to fidelity is the ultimate goal, researchers disagree about how much adaptation is allowed without compromising the integrity of an intervention (e.g., De Fazio, Fain, & Duchaine, 2011; Harn et al., 2013; Ogden & Fixen, 2014). Those who

privilege fidelity over adaptation contend that implementation should occur as intended by developers, whereas those who privilege adaptation over fidelity may more readily allow for changes to occur to fit specific contexts.

One example of researchers taking a stricter stance on fidelity is Elliott and Mihalic (2004). Elliott and Mihalic reported the findings from a major dissemination and replication project on violence prevention. Within their report on intervention fidelity, the authors voiced strong concerns against local adaptation for several reasons. First, they asserted that while fidelity requires only the implementation of core components as designed and demonstrated in trials, very few to no intervention programs have conducted a core component analysis to establish which components are core. Second, Elliott and Mihalic stated that a number of the key assumptions in the balanced approach to fidelity/adaptation are questionable, particularly that the local environment is an unchangeable given. Their experience suggested otherwise, and they contended the critical question may not be, "Will this program fit in this local context?" but, "How does this context have to change for us to successfully implement this program here?" (p. 50).

Elliott and Mihalic also questioned another assumption made by those who favor adaptation in that the only way to get local buy-in is to negotiate control over program content and the delivery process. Elliott and Mihalic reported this was clearly not the case in their study, and they were able to establish buy in through capacity building efforts; this included a local needs assessment across study sites and selection of appropriate programs, linkages, resources and local champions for the program. Finally, the authors discussed the assumption that local adaptation is inevitable. They stated they experienced very little local adaptation in their violence prevention initiative and were

able to achieve a high level of fidelity. Elliott and Mihalic concluded that the call for a negotiated balance in fidelity/adaptation has the potential for lowering what they term the *Gold Standard* of research, encouraging and empowering local implementers to make questionable adaptations, and undermining the research community's commitment to fidelity.

Researchers who take more of an adaptation perspective assert adaptation is inevitable (Lendrum & Humphrey, 2012; Maynard, Peters, Vaughn, & Sarteschi, 2013; Nelson et al., 2012; Sanetti & Kratochwill, 2012). Durlak & DuPre(2008) fall into this category and, in a review of five meta-analyses on factors affecting implementation, concluded that expecting perfect or near-perfect implementation is unrealistic. They claimed positive outcome results have often been obtained with levels of implementation around 60%, and few studies have attained levels greater than 80%. Durlak and DuPre also discussed the phenomena of implementation threshold effects that have occurred within certain studies. That is, although it might seem that ''more is always better," it is possible that once a certain level of implementation is attained (e.g., in dosage or fidelity), higher levels may not always lead to significantly better outcomes, particularly if the intervention's core components have been effectively delivered.

A number of researchers strongly advocate the need for local adaptation in order to match interventions to local conditions and, hence, believe flexibility with fidelity can actually improve outcomes (e.g., Castro, Barrera, & Martinez 2004; Century et al., 2011; Kendall, Gosch, Furr, & Sood, 2008). The need to adapt for individuals from different cultures is one such example and has shown promising results. One study demonstrating this promise involved the cultural adaptation of an evidence-based parent training

program for Spanish-speaking Latino parents. Domenech Rodriguez, Baumann, & Schwartz (2011) documented and detailed how adaptations to both process (e.g., engaging community leaders and parents, decentering of the program manual) and content (e.g., changing the appropriateness of the language, metaphors, and contexts to match the target population) of the parent training program were made in a very carefully planned a priori process involving a pilot study, focus groups, and testing of the adapted intervention. The authors reported that the positive outcomes of their study provide support for the idea that cultural adaptations can improve service delivery to diverse groups with a reasonable amount of work conducted a priori to implementation and can be conducted systematically with documentation for replication purposes. Positive outcomes in the study were reported via improved retention rate data, continued requests for services after data collected had stopped, and preliminary outcome data from intervention impact.

On a larger scale, Griner and Smith (2006) set out to determine if there was evidence that cultural adaptations are effective. Griner and Smith conducted a meta-analysis of nearly 80 studies that contained explicit statements that intervention content, format or delivery was adapted based on culture, ethnicity, or race. Their study produced a weighted average effect size of $d = 0.45$, indicating that culturally adopted interventions were moderately effective. It was noted, however, that these results varied depending on whether participants were of the same race ($d = 0.49$) or mixed race participants ($d = 0.12$). In sum, current evidence regarding the effectiveness of cultural adaptations to evidence-based practices and programs appears to be mixed, but promising.

In general, most other researchers hold a middle ground fidelity position, including Zvoch (2009), who stated that invariantly requiring strict fidelity to a program model or allowing widespread adaptation of key intervention components is likely to be counterproductive. He suggested that a more efficacious approach to program delivery would likely involve a strategic alignment between the treatment context, the specific aspects of the treatment intervention, the skills of the treatment provider, and the unique needs of the treatment recipient. Identifying the individual and contextual factors that promote or inhibit program adherence is thus one step in elucidating the conditions under which a specific deviation from protocol is likely to confer a clear advantage or, alternatively, undermine an otherwise effective treatment routine. Similarly, Lendrum, (2010) stated that to avoid over-modification and a resulting lack of impact, the emphasis should be on finding the right balance between fidelity and adaptation.

In sum, researchers in implementation science that have documented the process of adaptation note that the key to successful adaptation is when teachers understand and implement the "core" or essential components of the practice (Odom, 2009). The greater the number of modifications, the higher the risk that critical components might be changed, resulting in loss of impact on outcomes. However, some adaptation is inevitable, and developing a plan a priori for flexibility and fit is needed to ensure these adaptations do not compromise targeted outcomes.

Implementation Fidelity and MTSS

Methods of collecting fidelity data and MTSS. Tracking fidelity of implementation of multi-tiered systems, particularly in studies of RTI systems, is one of the most important components necessary to maximize program effectiveness (Keller-

Margulis, 2012; Kovaleski, 2007; Shinn, 2007). Fidelity of implementation becomes especially critical when multi-level intervention and data collection are used for potentially high-stakes decision making. In reality, however, fidelity of implementation has received relatively little attention within the RTI literature (Keller-Margulis, 2012; Noell & Gansle, 2007).

A number of tools have been developed to monitor the implementation of RTI at the school level. However, the technical properties of most of these instruments do not appear to be reported. One example is the *RTI Essential Components Integrity Rubric* and the *RTI Essential Components Integrity Worksheet* developed by the National Center on Response to Intervention. This instrument is a school level self-appraisal of RTI in which those individuals responsible for implementation score fidelity to various components of an RTI system. These elements are very similar to those components described within a Schoolwide Reading Model and are scored on a Likert-type scale of 1-5. Although comprehensive in nature, the instrument's technical properties including validity and reliability are not reported.

The Colorado Department of Education (2011) created a robust set of RTI rubrics at the district, school and classroom levels. Implementers are asked to rate implementation of RTI in the following general areas: leadership, problem solving, curriculum and instruction, assessment, positive school climate and family and community partnering. Rubrics were developed for each of these six areas and have four growth stages descriptions (Emerging, Developing, Operationalizing, and Optimizing). In using the rubrics, a school team self-reflects and discusses each of the stages for each of the general areas. The team then determines at what stage the school is currently

functioning in each of the areas. Although the process for creating the rubrics is described in much detail and the process for using the rubrics is very comprehensive, technical adequacy of the measures is not addressed.

Finally, the *Planning and Evaluation Tool for Effective School-wide Reading Programs-Revised* (PET-R; Kame'enui & Simmons, 2003) is a 38-item tool developed for evaluating the implementation of multi-tiered reading programs and is aligned with the Schoolwide Reading Model. School personnel self-report whether each of the items within the PET-R are fully in place (2 points), partially in place (1 point), or not in place (0 points). The instrument is described as a planning and evaluation tool. As with the tools previously described, the instrument has not been evaluated for validity and reliability.

Most recently, there does appear to be a growing effort to develop RTI implementation integrity tools to fill in the missing gap between the need for validated instruments and what currently is available. The *School Implementation Scale* (Erickson, Noonan, & Jenson, 2012) is one such tool. This instrument was developed through a multi-year iterative design process using selected schools within one state and conceptualized as a measure of school-wide implementation of one integrated academic/behavior RTI model. The authors reported the preliminary results of a recent pilot study of the psychometric properties of the *School Implementation Scale* as being highly reliable and providing valid data on the implementation of integrated academic/behavior RTI models within schools. Future plans call for expansion of implementation of the scale within other schools as well as states using other RTI models.

Noltemeyer, Boone and Sansosti (2014) also reported on a preliminary study using the *RTI Implementation Scale for Reading* (RTIS-R). Results suggested that the instrument is rigorous, has strong reliability and has potential for future use, although the authors reported more work is needed on its development. Finally, researchers at Florida's Multi-Tiered System of Supports (Stockslager, Castillo, Brundage, Childs & Romer, 2014) reported on the development of the *Self-Assessment of MTSS* (SAM), a self-assessment fidelity instrument for MTSS implementation. Their pilot study examined the psychometric properties of content validity, construct validity, reliability and predictive validity. Promising results were indicated, and future plans called for a much deeper look of an exploratory factor analysis review and an upcoming national validation study.

MTSS/RTI fidelity and student outcomes. As previously indicated, the relationship between implementation integrity of multi-tiered systems of support and student outcomes appears to be an area of needed research (Denton, 2012; Keller-Margulis, 2012). Harms (2010) postulated that one of the reasons for this lack of research is that many researchers only examine outcome data once a practice is fully implemented, while other studies describe implementation levels and then separately describe student outcomes, but never draw a connection between the two. In an attempt to close this research gap, Harms conducted a study to determine the extent to which outcomes in reading and behavior were associated with scores on implementation checklists. Using CBM-type reading scores, Harms reported the Pearson correlation between outcomes and fidelity as .135 ($p <$. 01). While the correlation was found to be statistically significant, the magnitude of the correlation was low. Interestingly, the

Pearson correlation reported between a fidelity measure of Positive Behavior Supports and Interventions and student outcomes for behavior was -.136 ($p < .01$). While the negative correlation was found to be statistically significant, again, the magnitude of this correlation was also low.

Relatedly, Mellard, Frey and Wood (2012) stated that although the framework of RTI has been widely accepted and adopted around the United States, the evidential validity of RTI has not yet been established. As a result, they also attempted to close this research gap with a study measuring, analyzing and reporting schoolwide student reading effects of RTI. In essence, a set of open-ended survey and interview items related to implementation of RTI components were provided to 60 schools in 16 states identified as using RTI, and staffs were given an opportunity to describe and document their RTI design features. Forty-one schools responded to the survey, and, using experts in the field to study the school responses along with a specific set of criteria, five schools were chosen for inclusion in the study as they were judged to be "sufficiently and verifiably implementing RTI components" (p. 28). Outcome measures in this study consisted of screening and progress monitoring measures that were in place at the five schools. Mellard et al. calculated effect sizes by adopting Shapiro and Clemens' (2009) proposed conceptual model for evaluating RTI system effects by comparing rates of improvement for each school to a national normative data set. In other words, they compared the effect sizes achieved in one year of instruction in the school's RTI system with the normal growth or effect size for each grade level.

Reported effect sizes differed by school and by grade level. The researchers summarized their findings from the five schools by stating that three types of results

emerged from the data. The first result they described as *accumulated advantage* in that the three schools with students generally scoring above the norm on their fall (baseline) tests not only maintained this advantage but gained more than expected during the year of tiered instruction. The second type of outcome was demonstrated by one school in which students began the year averaging well below normal and made substantial gains that closed the gap between these students and the test's normative sample. The third result was less positive. For one school, the average fall test scores were often above the norm, but students did not maintain this advantage in spring test scores. The authors hypothesized that this school scored lower possibly due to inferior general education instruction as determined by the experts' lower rating of fidelity of implementation to the school's core reading program.

In one additional study on MTSS fidelity and student outcomes, Parisi (2009) focused on implementation of one component of the RTI system – the relationship between the fidelity of implementation of research-based interventions and several kindergarten literacy measures. Fidelity was measured using direct observations of intervention teachers and through the use of corresponding fidelity checklists for each intervention. Specifically, among other purposes, the study looked at how dimensions of fidelity relate to student literacy outcomes using multi-level models. It was hypothesized that those teachers that had the highest fidelity scores over time would have students that performed the highest at the end of the intervention.

Unexpectedly, the relationship between average total fidelity and student outcomes was in the opposite direction of what was hypothesized. That is, lower total fidelity was related to higher student outcomes. It is noted that Parisi described important

limitations to the study including limited variability in student outcomes and significant issues in the methods for conducting fidelity observations.

The Current Study

From the research that is available on the relationship between implementation of schoolwide RTI systems and student outcomes, data up to this point appears sparse. Those studies that have been conducted appear to have mixed results regarding whether implementation predicts outcomes. The goal of the present study therefore is to add to the literature base in better understanding fidelity of implementation as it relates to student reading outcomes within MTSS. Specific questions for the study will be presented in the following chapter.

CHAPTER III

METHODS

The purpose of the current study was to determine the relationship between fidelity of implementation measures used during Reading First and outcome measures for students within each of these Reading First schools in Oregon. Extant data from the previously completed Oregon Reading First program was used to answer questions within the study. While student outcome data was collected every school year within the project, school implementation data used in this study was collected and recorded only during the second year of implementation – during the 2004-05 school year for Reading First Cohort A schools. Hence, this study looks at data collected during the 2004-05 school year.

This relationship was evaluated using both direct and indirect measures of program fidelity. More specifically, these independent measures included three indices of implementation fidelity created from data collected during Oregon's implementation of the Reading First program, as well as a combination of all three indices. These three indices, as described more thoroughly below, included (a) an *Oregon Reading First (ORF) Implementation Compliance Index*, which included a record of submissions of required project deliverables that documented implementation of the major components of the Schoolwide Reading Model as well as limited observation data; (b) the *Professional Development Attendance Record,* which included attendance records of teachers, principals, reading coaches, and district team members at required Reading First professional development functions, district leadership sessions, and coaches' meetings; and, (c) the *Continuation Application Index* that, in essence, was a school's self-report of

implementation of the Schoolwide Reading Model and was evaluated by Reading First Program staff. Additionally, a *Total Composite Score* of all three measures was utilized as a fourth overall measure of implementation fidelity. All of the fidelity measures used within the present study were aligned with the concept of fidelity compliance or adherence to the Oregon Reading First Schoolwide Reading Model rather than process or quality. Hence, the term *Implementation Compliance* will be used hereafter to describe these measures. Dependent or outcome measures within the present study included three measures of oral reading fluency collected across one school year for each student, as well as one reading comprehension measure for each student in second and third grades. The specific research questions related to the present study were as follows:

1. To what extent is school-level variance in student growth on curriculum based measures of oral reading fluency in second and third grades in Oregon Reading First schools associated with higher levels of implementation as measured by a composite index of three indices of implementation compliance?
2. To what extent does each of the three indices of fidelity to the Schoolwide Reading Model independently explain school-level variance in student growth on curriculum-based measures of oral reading fluency in second and third grade in Oregon Reading First schools?
3. To what extent is school-level variance on the Stanford Achievement Test-10 (second grade) and Oregon Test of Knowledge (third grade) in Oregon Reading First schools associated with higher levels of implementation of the Schoolwide Reading Model as measured by a composite index of three indices of implementation compliance?

4. To what extent do the three different indices of implementation compliance independently explain student performance on summative, end-of-year reading achievement measures of the Stanford Achievement Test-10 and Oregon Test of Knowledge and Skills?

Given the fact that, in general, higher degrees implementation are associated with higher outcomes, it was hypothesized that greater levels of implementation of the Schoolwide Reading Model within Reading First schools would be associated with greater student growth in oral reading fluency for second and third grade students within those schools across all three compliance indices. Additionally, it was hypothesized that higher levels of fidelity to the Schoolwide Reading Model by Reading First schools would also be associated with greater comprehension performance for students within these schools. In particular, it was theorized the *Composite Point Index* would be particularly significantly predictive of student reading outcomes as this measure was a combination of several different methods of measuring implementation compliance.

Study Participants

Data for this study were obtained from 5,283 second and third grade students within the first cohort (known as Cohort A) of 34 Oregon schools participating in Reading First during the 2004-2005 school year. Demographic characteristics of this group can be found in Table 1. As noted, over half of the students in the study were eligible for Free and Reduced Lunch (FRL) (although FRL data was missing for 12.5% of the students in second grade and 8.4% for students in third grade). A majority of the students were of a minority status, with nearly one-third of all students of Hispanic

Table 1

Demographic Information of Study Participants

	Grade 2 *n* = 2,653		Grade 3 *n* = 2,630	
Variable	*N*	%	*N*	%
Females	1270	48.6	1198	45.6
FRL	1366	51.5	1781	67.7
SPED	327	12.3	336	12.8
LEP	904	34.1	815	31.0
White	1218	45.9	1202	45.7
Hispanic	844	31.8	814	31.0
Black	238	9.0	243	9.2

Note. FSL = free or reduced lunch eligibility; SPED = students identified for special education; LEP = Limited English proficiency.

background. Relatedly, over one-third of the students in both second and third grades were students classified as having limited English proficiency.

Measures

Dependent Variables

Data from three measures were collected and served as dependent variables for the current analysis. These included the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) 6th Edition Oral Reading Fluency assessment (DORF) (Good & Kaminski, 2002), the Stanford Achievement Test 10th Edition (SAT-10) (Harcourt Assessment, Inc., 2004), and the Oregon Assessment of Knowledge and Skills (OAKS) Reading/Literature (Oregon Department of Education, 2004). The technical characteristics of each assessment are reported below. The DORF was administered three times during the 2004-05 school year to students in both second and third grades. The SAT-10 was administered to all second grade students one time during the 2004-05

school year, and the OAKS Reading/Literature assessment was administered to all third grade students during SY 2004-05 to obtain a measure of overall reading achievement.

DIBELS. The DIBELS Oral Reading Fluency (DORF) assessment was the sole DIBELS measure used within this study. DORF is a standardized, individually administered test of accuracy and fluency of reading connected text. Student performance is measured by having students read aloud three different passages for one minute, and the median number of words read correctly per minute is recorded as the oral reading fluency score. This procedure occurred three times per year at the beginning, middle, and end of the school year. Test-retest reliabilities for second and third grade Oral Reading Fluency range from .92 to .97, and alternate-form reliability from the same test level range from .97 to .99 (Baker et al., 2008; DMG, 2007). Concurrent validity estimates with the SAT-10 report correlations of .67 at the second grade level and .80 at the third grade level (Baker, et al., 2008). McKenna and Good (2003) reported a concurrent validity coefficient of .69 with the Oregon State Assessment.

SAT-10. The SAT-10 is a multiple-choice, standardized assessment and was administered to all second grade students within Oregon Reading First schools to assess overall reading proficiency. The SAT-10 is a norm-referenced, group-administered assessment and was given at the end of the 2004-05 school year by classroom teachers. Test administration time varies per publisher instructions depending upon grade level and subtests given. The total Standard Score, based on grade, was used in all analyses; subtests administered in Grade 2 were *Word Study Skills*, *Reading Vocabulary*, and *Reading Comprehension*. The Kuder-Richardson reliability coefficient for the total

reading score of the SAT-10 at Grade 2 is .95, while validity coefficients between the total reading score and the *Otis-Lennon School Ability Test* range from .61 to .74.

OAKS Reading/Literature. The Oregon Test of Knowledge and Skills in reading (OAKS) is a statewide assessment given to all students in Oregon starting in third grade to assess overall reading performance. During the 2004-2005 school year, students were allowed to take the test either by pencil-and-paper or online with a testing time of approximately 120 minutes. Students had the opportunity to take the assessment three times during the school year with testing dates scheduled from September through May, and the highest score was reported for accountability purposes.

The OAKS Reading/Literature Test is a multiple-choice test with approximately 50 items with results recorded in four score reporting categories. The categories associated with knowledge and skills typically required for reading comprehension include: (a) vocabulary knowledge, (b) reading to perform a task, (c) demonstrating general understanding, and (d) developing an interpretation. Students receive a raw score on the OAKS, which is converted into a scaled Rasch Unit score (RIT score) based on the number of questions answered correctly compared to the total number of questions on the form, while taking into account the difficulty level of the questions. A higher RIT score indicates a higher level of achievement.

Concurrent validity with the California Achievement Test (1992), Iowa Test of Basic Skills (1998), NWEA Subject Tests (2003-2004) and Lexile Scale (2004) ranged from .73 to .78. Four alternate forms of the third grade reading test yielded an internal consistency estimate of .95. The alternate format correlation coefficients between pencil-

and-paper and computer-based item calibrations were high, with the third grade coefficient reported as .96.

Independent Variables

Implementation data was collected through three different indices of fidelity. These indices have been renamed as follows: (a) the ORF Implementation Compliance Index; (b) the Continuation Application Index, and (c) the Professional Development Attendance Record. In addition, a combination of point totals from all three indices of fidelity was used in this study to create a fourth independent variable termed the Total Composite Point Index. Each of these indices is described below.

Oregon Reading First (ORF) Implementation Compliance Index. *ORF Implementation Compliance Index* is the title given to a set of required documents schools were required to develop and submit for documenting and demonstrating compliance with Oregon's Reading First implementation plan. The major required documents of the ORF Implementation Index along with a description of each are listed in Table 2. A score was assigned to each school based on complete submission and timeliness of required documents throughout the 2004-05 school year. The *Implementation Compliance Index* can be considered as an adherence tool rather than a quality or process fidelity tool as in general, schools were awarded points simply for submitting documents in a timely manner, rather than points being assigned for quality of implementation. As noted, a total of 35 points was the maximum score possible with this index.

Continuation Application Index. Federal Reading First guidance stated that continuation awards to local educational agencies must take into account the

Table 2

Components of Implementation Compliance Index

Component	Description	Point(s) Possible
Planning and Evaluation Tool for Effective Schoolwide Reading Programs – Revised (PET-R) (Kame'enui & Simmons, 2003)	Self-assessment tool in which Reading First school sites measure level of implementation of the major components of the Schoolwide Reading Model (SWRM). One point for completion and submission of the PET-R	1
Reading Action Plan	School-level plan for carrying out needed action(s) to improve components of the SWRM. One point awarded for completion of Reading Action Plan .	1
Fidelity Observation 1 (Fall)	Observation conducted by Reading First Coach and Oregon Reading First mentor to determine if classroom activities align with School-Level Reading First Plan. One point awarded at each grade level for compliance.	4
Fidelity Observation 2 (Winter)	Observation conducted by Reading First Coach and Oregon Reading First mentor to determine if classroom activities align with School-Level Reading First Plan. One point awarded at each grade level for compliance.	4
Fidelity Observation 3 (Spring)	Observation conducted by Reading First Coach and Oregon Reading First mentor to determine if classroom activities align with School-Level Reading First Plan. One point awarded at each grade level for compliance.	4
Core, Supplemental and Intensive (CSI) Instruction Map – Fall	Plan mapping out how each group of students (Core, Supplemental, Intensive) at each grade level would receive appropriate instruction. One point awarded per grade level .	4
Core, Supplemental and Intensive (CSI) Instruction Map – Winter	Plan mapping out how each group of students (Core, Supplemental, Intensive) at each grade level would receive appropriate instruction. One point awarded per grade level.	4
School Profile	School profile submitted including demographics and current reading data for all students at each grade level.	1

SAT-10 Observation	SAT-10 observation by Reading Coach indicating standardized administration procedures were followed. Point awarded for adherence to standardized procedures.	1
Grade Level Reading Action Plans (RAPs)	Each grade level completed action planning process and included a review of grade level data, identification of systems that need support, and creation a plan for change. One point awarded per grade level for completion.	4
Professional Development Needs Assessment	School conducted needs assessment, and professional development plan aligned with identified needs.	3
Lesson Progress Reports (LRP)	Summary submitted of the number of lessons taught in a defined time period, theme skill and in-program test results, and reading group members. One point awarded per grade level for completion.	4
Total Points Possible		35

progress made in improving reading achievement and implementation of its Reading First program as defined in its subgrant application; this guidance further stated that funding may be discontinued to any local education agency that was not making substantial progress. As a result, each school within Oregon Reading First was required to submit a Reading First Continuation Application at the end at the end of Year 2 of implementation (SY 2004-05).

The Oregon Reading First Continuation Application consisted of schools providing information on current implementation of their Reading First program in the following five sections: (a) summary and analysis of student performance, (b) fidelity of implementation, (c) leadership, (d) district support, and (e) budget. A series of questions and activities were required for schools to answer and complete within each section that were then scored on a 5-point rubric. The Continuation Application in its entirety can be found in Appendix A.

The Continuation Applications were then scored by two trained raters according to a scoring rubric developed by the Oregon Reading First Center and Reading First team at the Oregon Department of Education. If the two raters disagreed on any item, a third rater mediated the discrepancy by determining a rating with the benefit of access to the previous raters' rationales. All identifiable information was omitted from the reports before they were assigned to raters. A total score of 25 points was possible for the Continuation Application.

Professional Development Attendance Record. The third fidelity measure reflects the attendance of teachers, reading coaches, building principals, and district team members at required Reading First professional development functions focused on the Schoolwide Reading Model, coaches' meetings, and principal/district leadership sessions. A total of 26 total points was possible for attendance at required events. Attendance was monitored and recorded by State Reading First personnel via sign-in sheets at each event and permanently recorded on an Excel spreadsheet.

Total Composite Point Index. This measure was a composite point total, calculated as the simple sum of the three measures previously described. The point total included the *ORF Implementation Compliance Index* (35 points), the *Continuation Application Index* (25 points) and the *Professional Development Attendance Record* (26 points). Total possible point value for the *Composite Point Total* was 86 points.

Procedures

The questions asked within this study had a multilevel structure. For Questions 1 and 2, three oral reading fluency scores at three time points across the school year (Level 1) were nested within each of 4,485 students (Level 2) at Grades 2 and 3. These students

(Level 2) were nested within 34 schools (Level 3). For Questions 3 and 4, two measures (one for second grade students and one for third grade students) of overall reading proficiency, including comprehension, for each student (Level 1) were nested within 34 schools (Level 2).

Because of the nested makeup of the data, and because the most appropriate methodology for measuring changes in student achievement is through estimation of individual growth trajectories by means of the multilevel model (e.g., Bryk & Raudenbush, 1992, Zvoch & Stevens, 2003), Hierarchical Linear Modeling (Bryk & Raudenbush, 1992) was used to analyze the data. Additionally, the Hierarchical Linear Modeling (HLM) program, version 7.01 (Raudenbush, Bryk, Cheong, & Congdon, 2013) was used to estimate the two and three-level longitudinal models.

Questions 1 and 2 were tested using a three-level HLM model. Level 1 was a longitudinal growth model that fit a linear regression function to each individual student's DIBELS achievement scores over SY 2004-05. Equation 1 depicts this Level 1 model:

$$(1) \qquad Y_{tij} = \pi_{oij} + \pi_{1ij}\,(\text{Time}) + e_{tij}$$

Within the Level 1 model, Y_{tij} is the reading outcome (oral reading fluency score) at time *t* for student *i* in school *j*, π_{0ij} is the initial status of student *i* in school *j*, and π_{1ij} is the linear growth across the school year.

Level 2 Model:

$$(2) \qquad \pi_{0ij} = \beta_{00j} + r_{0ij}$$
$$\pi_{1ij} = \beta_{10j} + r_{1ij}$$

Level 3 Model:

$$(3) \qquad \beta_{00j} = \gamma_{000} + \gamma_{001}(W_{1j}) + u_{00j}$$
$$\beta_{10j} = \gamma_{100} + \gamma_{101}\,(W_{1j}) + u_{10j}$$

Mixed Model:

$$\gamma_{ijk} = \gamma_{000} + \gamma_{001}(W_{1j}) + u_{00j} + \gamma_{100}\,(\text{Time}) + \gamma_{101}(W_{1j})(\text{Time}) + u_{10j} + r_{0ij} + r_{1ij} + e_{tij} \quad (4)$$

For Question 2, each fidelity index was tested independently. Additionally, if multicollinearity was indicated during the examination of SPSS correlation statistics, this issue was to be further examined using all three indices of fidelity within a separate model.

Question 3 used the combined index of fidelity of implementation as the Level 2 predictor of *SAT10* and the *OAKS Reading/Literature* outcomes. For Question 4, each of the three indices of implementation compliance separately served as Level 2 predictors of outcomes on the *SAT-10* and the *OAKS Reading/Literature* assessments. In both questions, students nested within schools served as the two-level model. The mixed model used for Questions 3 and Question 4 is depicted in Equation 5:

$$Y_{ij}\,(Score) = Y_{oo} + Y_{01}Index_j + u_{0j} + u_{1j} + r_{ij} \quad (5)$$

CHAPTER IV

DATA ANALYSIS AND RESULTS

Data Analysis

The initial section discusses the actions taken to ensure the integrity and reliability of the data used for analysis. Missing data patterns and the steps taken to understand these patterns are initially examined. Descriptive statistics for all study variables were also studied to evaluate normal distribution, skew, and univariate outliers using SPSS Version 22 (IBM Corporation, 2013). Additionally, correlations among the three predictor variables of implementation integrity were examined and are discussed.

Missing Data

Data were first analyzed for missingness using the SPSS Missing Values Analysis (MVA). The chi-square statistic used for testing whether values were missing completely at random was Little's Missing Completely at Random (MCAR) test (Hill, 1998). Little's MCAR test for outcome data across second and third grades resulted in a chi-square of 132.80 ($df = 29$; $p < .001$), and a chi-square of 130.90 ($df = 31$; $p < .001$), respectively, which indicated that data was not missing at random at both grade levels. Given the fact that MNAR is not ignorable and can lead to biased interpretations (Behi, Goodson, & Neilands, 2008; Peugh & Enders, 2004), a number of procedures were utilized to better understand this MNAR issue.

The extent of missingness for DIBELS ORF for second and third graders during the 2004-05 school year is depicted in Table 3. Missing data across the year ranged from 9.2% to 12.0% across the fall, winter and spring benchmarking periods for ORF in Grades 2 and 3. Also noted is the fact that a significant number of student scores

(approximately 23%) were missing from the end-of-year overall reading measurements from the SAT-10 (second grade) and OAKS-Reading (third grade).

Table 3

DIBELS ORF, SAT-10, OAKS-Reading Missing Data SY 2004-05

	Grade 2			Grade 3		
	Valid	Missing		Valid	Missing	
	N	*N*	%	*N*	*N*	%
Oral Reading Fluency-Fall (ORF-F)	2372	281	10.6	2367	263	10.0
Oral Reading Fluency-Winter (ORF-W)	2392	261	9.8	2387	243	9.2
Oral Reading Fluency-Spring (ORF-S)	2341	312	11.8	2314	316	12.0
Stanford Achievement Test-10 (SAT-10)	2051	602	22.7	--	--	--
Oregon Assessment of Knowledge and Skills-Reading (OAKS-Reading)	--	--	--	2032	598	22.7

Note. Second grade students were not administered the SAT-10 and third grade students were not administered the OAKS-Reading.

Initially, to understand more precisely which students this missing data represented, a cross-tabulation of all outcome measures with student demographics was conducted and summarized in Table 4. Several statistically significant differences in missingness were noted. In both summative measures, statistically significant differences

Table 4

Percent of Missingness by Demographics Across Grades 2 and 3

	DIBELS-F	DIBELS-W	DIBELS-S	SAT-10	OAKS-Reading
Race					
Hispanic	10.94	9.38	10.31	29.01*	24.92*
American Indian/Alaska Native	9.09	7.27	9.09	16.98*	16.07*
Asian/Pacific Islander	8.93	8.93	8.93	16.07*	14.29*
Black	12.90	9.68	8.60	23.08*	18.95*
White	8.97	8.33	11.11	17.52*	20.51*
Multiple	12.50	12.50	12.50	28.57*	22.22*
Special Education					
Yes	10.29	7.20*	8.00*	22.76	24.22
No	10.40	9.94*	12.34*	22.69	22.48
Limited English Proficiency					
Yes	9.23	6.77**	7.08**	23.17	22.26
No	10.81	10.81**	14.07**	22.46	22.90
Free/Reduced Lunch					
Yes	11.13**	7.37**	7.67**	21.26*	16.78*
No	7.16**	4.48**	4.18**	12.62*	12.64*

*p = <.05, ** p = < .001

in missingness were determined for Race in Grade 2 ($\chi^2 = 45.14$, $p < .001$) and Grade 3 ($\chi^2 = 13.65$, $p < .001$). Students of Hispanic background had the highest rate of missingness compared to other races with 29% of student scores missing in the SAT-10 assessment at Grade 2 and nearly 25% of student scores missing on the OAKS-Reading at Grade 3. Asian students, on the other hand, showed the least amount of missingness with approximately 16% and 14% of students missing scores on these two assessments, respectively. Statistically significant differences were also found between special

education students and students without disabilities during the Winter ($\chi^2 = 5.95$, $p < .001$) and Spring ($\chi^2 = 9.34$, $p < .001$) DIBELS benchmarking periods, with special education students missing less scores than non-special education students. Similar results were found for Limited English Proficiency students. LEP students had statistically significant less missingness than non-LEP students in the Winter ($\chi^2 = 21.14$, $p < .05$) and Spring ($\chi^2 = 53.14$, $p < .05$) benchmarking time periods. Finally, statistically significant differences in missingness were found between students with Free/Reduced Lunch status and those without the designation for all assessments used in the study. Students with the FRL status experienced more missingness during the Fall ($\chi^2 = 19.14$, $p < .05$), Winter ($\chi^2 = 14.16$, $p < .001$), and Spring ($\chi^2 = 20.39$, $p < .05$) DIBELS assessments, as well as the SAT-10 ($\chi^2 = 28.81$, $p < .001$) in Grade 2 and OAKS-Reading ($\chi^2 = 6.29$, $p < .001$) in Grade 3.

Next, missing data was examined at the school level by cross-tabulating frequencies of missing patterns across all schools by grade level and then by schools. Table 5 reveals patterns of missingness for ORF across the 2004-05 school year by grade level. Similar patterns were found across second and third grades. A majority of students in the study had scores for all four outcome scores. As noted, the most frequent pattern of missing data was one in which students had all three DIBELS measures across the year but were missing outcome measures of either SAT-10 at second grade or OAKS-Reading data at third grade. Other less-frequent errors patterns appeared mixed.

An examination of the amount as well as patterns of missing data by school was then conducted through cross-tabulations using missing versus expected counts and percentages. Table B1 and Table B2 found in Appendix B summarize this information

Table 5

Most Common Patterns of Missingness DIBELS ORF, SAT-10 and OAKS by Grade Level for School Year 2004-2005

	Grade 2					Grade 3				
Sample Patterns	*N*	D-F	D-W	D-S	SAT-10	*N*	D-F	D-W	D-S	OAKS
	1873	×	×	×	×	1827	×	×	×	×
	211	×	×	×	O	241	×	×	×	O
	160	×	O	O	O	149	×	O	O	O
	122	O	×	×	×	125	O	×	×	×
	101	×	×	O	O	113	×	×	O	O

Note. × = Score, O = No Score

across all four outcome measures at each grade level. Differences in expected versus actual missingness in Grade 2 reached up to 12.8%, while expected versus actual missingness in Grade 3 reached up to 11.6% in one school. A cross-tabulation of missing data by schools and all outcomes indicated that missing data was not distributed at random. For Grade 2, statistically significant differences on missingness between schools were noted on Fall ($\chi^2 = 65.38$, $p < .001$), Winter ($\chi^2 = 57.70$, $p < .05$) and Spring ($\chi^2 = 50.35$, $p < .05$) DIBELS measures. In regard to Grade 3, statistically significant differences in missing data between schools were found for Winter ($\chi^2 = 57.80$, $p < .05$) and Spring ($\chi^2 = 59.08$, $p < .05$) DIBELS scores. Additionally, statistically significant differences in missingness across schools were noted for the summative measures of SAT-10 ($\chi^2 = 60.61$, $p < .001$) and OAKS-Reading ($\chi^2 = 80.90$, $p < .001$) at Grades 2 and 3, respectively.

To further examine the MNAR result, a two-way analysis of variance between all outcome measures and all potential data patterns were conducted by measure and found to be statistically significant ($\chi 2 = 818.85$, $p < .001$). A comparison of estimated marginal outcome means along with associated standard errors for missing data patterns in reported in Table B3 in Appendix B. An examination of post-hoc Bonferroni corrections for these analyses showed statistically significant results for a number of these comparisons as shown in Table B4 in Appendix B. Of particular note, students with the pattern of scores on all four measures (XXXX) were statistically significantly different than students with a majority of other missing patterns across various DIBELS time point measurements including both the SAT-10 and OAKS-Reading measures. For example, students in group pattern XXXX statistically significantly outperformed students with the pattern of XXXO in all DIBELS measures at grades 2 and 3.

Although the reasons for these statistically significant differences cannot be determined, hypotheses as to why lower performance did occur in students with missing data patterns include at-risk characteristics and academic difficulties associated with family and student mobility (Blazer, 2007; Grigg, 2012; Mehana & Reynolds, 2004; Thompson, 2011; Xu, Hannaway, & D'Souza, 2009), and lower performance associated with poor attendance (Gottfried, 2009). Another possible explanation is that schools may have excluded lower-performing students from taking the final assessment. Olson (2003) discussed issues of purposeful exclusion in NAEP assessments and found that increases in exclusion rates were correlated with increases in NAEP reading scores at the state level. No matter the potential causes of these differences in outcomes based upon data patterns, these results indicate that findings from this study will be biased due to the fact

that students with XXXX patterns were the higher performing students. Additionally, generalization of findings will be limited to this study and will be reflected in the discussion chapter.

Descriptive Statistics and Tenability of Statistical Assumptions

Descriptive statistics for all outcome measures are shown in Table 6. A visual inspection of histograms, Q-Q plots and box plots showed that, across second and third grades, ORF-W and ORF-S measures were approximately normally distributed for both grades. ORF-F scores were positively skewed for both second and third grades, with ORF-F second grade scores being particularly skewed. This would not be an unusual finding given the fact that reading fluency just starts to develop during this initial fall time period. Additionally, data sets for both the SAT-10 scores for second grade students and OAKS-Reading scores for third grade students were found have approximate normal distributions.

Table 6

Mean, Standard Deviation, Minimum and Maximum Scores for Outcome Scores SY 2004-05

	Grade 2					Grade 3				
	N	*M*	*SD*	Min	Max	N	*M*	*SD*	Min	Max
ORF-F	2372	37.29	30.16	0	201	2367	62.46	35.55	0	222
ORF-W	2392	63.21	38.65	0	216	2387	79.63	39.64	0	236
ORF-S	2341	80.32	40.12	0	234	2314	97.45	39.51	0	226
SAT-10	2051	584.65	43.05	455	732					
OAKS-Reading						2032	209.59	10.57	178	249

School level descriptive statistics related to implementation fidelity measures are depicted in Table 7. An examination of frequency distributions showed normality for both the Continuation Application Index as well as the Composite Score Index; negative skews were obtained for both the Professional Development Attendance Record and

Table 7

Mean, Standard Deviation, and Minimum and Maximum Scores for Implementation Fidelity Indices for 34 Oregon Reading First Schools

Index	*M*	*SD*	Min	Max
Implementation Compliance Index	29.6	6.3	7	35
Continuation Application Index	16.2	3.4	7	22
Professional Development Attendance Record	21.9	6.4	2	26
Composite Score Index	67.1	11.2	23	81

Implementation Compliance Index. Large ranges of scores were present for all four of the school level predictors. In particular, the Composite Score Index showed a range of scores from 23 points to 81 points, with a total possible index score of 86 points. It is notable that one school in particular, which will be called *School 44* within the present study, was a significant outlier with regard to several fidelity measures, including the Composite Point Index. To illustrate this point, Table 8 contains fidelity scores for School 44 compared to the remaining 33 schools. As a result of these differences, all HLM analyses were conducted with and without School 44 and will be further discussed within the results section. Finally, to test the assumption of independence between the three distinct measures of implementation fidelity, intercollelations between these measures were calculated using SPSS. Results produced tolerance levels ranging from

Table 8

Comparison of Average Fidelity Scores With School 44

	Implementation Compliance Index	Continuation Index	Professional Development Attendance Record	Composite Point Index
School 44	7.00	14.00	2.00	23.00
Other 33 Schools	30.24	16.21	22.42	68.87

.871 to .997, and variance inflation factors ranging from 1.02 to 1.14, providing evidence that multicollinearity was not an issue with these three measures.

Results

Results of the HLM analyses are presented below. Each research question is restated, and results are organized by research question. Research questions 1 and 2 are both three-level models and analyze the relationship between DIBELS ORF growth over the 2004-05 school year and measures of implementation compliance at the school level. Questions 3 and 4 use a two-level HLM model and examine the relationship between overall reading measures of the SAT-10 and OAKS-Reading at second and third grades, respectively, and previously-described school-level implementation compliance measures.

Questions 1 and 2 With Full Sample

The first series of models were targeted at answering research questions 1 and 2, which are reproduced below:

1. To what extent is school-level variance in student growth on curriculum based measures of oral reading fluency in second and third grades in Oregon Reading

First schools associated with higher levels of implementation as measured by a composite index of three indices of implementation compliance?

2. To what extent does each of the three indices of fidelity to the Schoolwide Reading Model independently explain school-level variance in student growth on curriculum-based measures of oral reading fluency in second and third grade in Oregon Reading First schools?

Results are reported for Grade 2 and then Grade 3 with School 44. The same results are then reported for Grades 2 and 3 without School 44. Reporting progresses from unconditional means and unconditional growth models, to the Composite Point Index model and then to HLM models using the three separate indices of fidelity. Additionally, although not a targeted question for the study, the role of school-level demographics as they relate to implementation compliance measures is also addressed.

Grade 2. In order to begin answering Questions 1 and 2, an unconditional one-way Analysis of Variance (ANOVA) was conducted to provide information about the total variance observed in DIBELS ORF scores as accounted for by each of the three levels of the model, as well as test the hypothesis that the variability is null. As noted in Table 9, the unconditional means model revealed an Intraclass Correlation Coefficient (ICC) of .019 indicating about 1.9% of the variance in DIBELS scores was accounted for between schools (Level 3). Results from the unconditional growth model showed a small increase of variance at Level 3, with a Level 3 ICC of .043. Although the variance between schools appears small, it was statistically significant. As noted by Roberts (2007), even with intraclass correlations near zero, group dependence can exist when

Table 9

Fixed Effects Estimates and Random Variance Estimates for Predictor Models on DIBELS Growth Grade 2

	Fixed effects					
Parameter	Unconditional Means	Unconditional Growth	Composite Index	Implementation Compliance Index	Continuation Index	Attendance Record
Intercept	58.91**	80.23**	80.24**	80.23**	80.20**	80.26**
Composite			0.21			
Compliance				0.29		
Continuation					0.72	
Attendance						0.12
For Time_2 slope						
Intercept		21.31**	21.30**	21.31**	21.30**	21.30**
Composite			0.14**			
Compliance				0.18*		
Continuation					0.20	
Attendance						0.14
	Random effects and model fit statistics					
Intercept, e	616.32	94.05	94.07	94.07	94.06	95.07
Intercept, r_0	998.41**	1542.40**	1542.19**	1542.41**	1542.24**	1542.25**
Time_2 slope, r_1		50.48**	50.43**	50.46**	50.47**	50.45**
Intercept, μ_{00}	31.48**	65.88**	60.28**	62.64**	60.53**	64.89**
Time_2, μ_{10}		9.63**	7.45**	8.45**	9.20**	8.88**
ICC Level 3	1.91%	4.3%				
Level 3 Pseudo R^2			10.30%	5.85%	7.65%	2.30%
Deviance	79,209.92	62,837.19	62,827.19	62,832.36	62,834.77	62,832.20
Parameters	4	9	11	11	11	11

*p = <.05, ** p = < .001

variables are added to the model. Hence, it was appropriate to include Level 3 predictors in this model.

Within the unconditional growth model and models that follow hereafter, time points were coded as -2 (beginning of the year), -1 (middle of the year), and 0 (end of year) so as to represent growth over the year as predicted by measures of fidelity. As a result, the intercept represents the mean DIBELS score at the end of the school year and is coded as time point 0. The unconditional growth model estimates indicated students were able to read approximately 80 words correct per minute on average at the end of the school year ($\gamma_{000} = 80.23$). Students gained an average of 21.31 additional words between each assessment time period (γ_{100}).

Information on the extent to which the Total Composite fidelity index predicted DIBELS outcomes is also depicted in Table 9. The Composite Point Index was a small, but statistically significant predictor of the DIBELS ORF growth slope ($\gamma_{101} = 0.14$, $p < .001$). By adding this predictor to the model, the Level 3 overall variance was reduced by 10.30%. Estimates indicated that students in a school with an average composite total were able to read approximately 80 words correct per minute average at the end of the school year ($\gamma_{000} = 80.24$), and gained an average of 21.30 words between each assessment time period. Students in a school with above an average composite point total made an additional 0.14 words per minute gain on the DIBELS slope for each point a school performed above the average, while students in schools with below average composite point totals experienced 0.14 words per minute less growth for each point their school was below the composite point total. To put this in perspective, students in a school that scored one standard deviation above the average school mean for the

Composite Point Index had additional average gains of approximately 1.57 words per minute on the DIBELS growth slope.

In order to estimate the variance in ORF growth associated with each separate school level measure of implementation compliance, each predictor was added to the model independently of the other predictors. This included the Implementation Compliance Index, Continuation Index, and Professional Development Attendance Record. Each predictor was added to the model as grand-mean centered across all 34 schools. Thus, the intercept was the predicted outcome value at the end of the year for a school with the average score on each single predictor.

All three predictors slightly lowered the variance at Level 3 with pseudo R^2 percentages ranging from 7.65% (Continuation Index) to 2.30% (Professional Development Record). As Table 13 indicates, of the three independent fidelity indices, only the Implementation Compliance Index was a statistically significant predictor of DIBELS scores. Specifically, the Implementation Compliance Index was predictive of the ORF DIBELS growth slope ($\gamma_{101} = 0.18$, $p < .05$). Thus, for students in a school with average performance on the compliance index, a one point difference on the compliance index was associated with a 0.18 per word difference in DIBELS growth rate favoring the school with higher implementation. This relates to a 1.13 per word difference on the ORF DIBELS growth for a school one standard deviation above or below a school with an average compliance index score.

Grade 3. Similar analyses were conducted using Grade 3 data in order to answer Questions 1 and 2; results are summarized in Table 10. An unconditional means one-way Analysis of Variance (ANOVA) produced a Level 3 ICC of .03, and an unconditional

growth model revealed a Level 3 ICC of .05. Thus, within the growth model approximately 5% of the total variance between DIBELS ORF scores was accounted for at Level 3, with the remaining 95% variance accounted for at Levels 1 and 2. Once again, although small, the variance at Level 3 was statistically significant.

As noted in Table 10, the Total Composite Point Index was a statistically significant predictor of spring DIBELS outcomes at Grade 3 ($\gamma_{101} = 0.34$, $p < .05$). These results indicate that for students in a school with a an average composite point total, a one point difference on the composite total was associated with a 0.34 per word difference in DIBELS spring scores, once again favoring the school with higher implementation. Cumulatively, this relates to a 3.80 per word difference on the ORF DIBELS spring score for a school one standard deviation above or below a school with an average compliance index score. Using the Total Composite Point index as the sole predictor in the model, the overall Level 3 variance was reduced by 20.04% compared to the unconditional growth model.

Each of the three single measures of implementation fidelity was then added to the model independently and compared against the unconditional growth model. Mixed results were found. Both the Continuation Index and Attendance Record failed to produce any statistically significant results. In contrast, the model that included the Implementation Compliance Index did result in a statistically significant predictor of spring DIBELS scores ($\gamma_{101} = 0.62$, $p < .05$) and produced a pseudo R^2 of 21.76%. Students in a school with an average score on this fidelity measure had a mean performance of 99.70 correct words per minute in spring of Grade 3. Students in a school

Table 10

Fixed Effects Estimates and Random Variance Estimates for Predictor Models on DIBELS Growth Grade 3

Parameter	Unconditional Means	Unconditional Growth	Total Composite Index	Implementation Compliance Index	Continuation Index	Attendance Record
Fixed effects						
Intercept	82.60**	99.74**	99.70**	99.78**	99.72**	99.72**
Composite			0.34*			
Compliance				0.62*		
Continuation					0.43	
Attendance						0.30
For Time_2 slope						
Intercept		17.89**	17.89**	17.89**	17.89**	17.89**
Composite			0.05			
Compliance				0.04		
Continuation					0.07	
Attendance						0.09
Random effects and model fit statistics						
Intercept, e	420.34	80.56	80.56	80.57	80.56	80.56
Intercept, r_0	1107.94**	1336.07**	1336.25**	1336.45**	1336.17**	1336.02**
Time_2 slope, r_1		22.00**	22.00**	22.00**	22.00**	22.00**
Intercept, μ_{00}	49.92*	68.00**	53.50**	51.95**	65.60**	64.68**
Time_2, μ_{10}		6.14**	5.78**	6.06**	6.09**	5.85**
ICC	3.16%	4.90%				
Level 3 Pseudo R^2			20.04%	21.76%	3.30%	4.87%
Deviance	56,200.10	50,759.28	50,753.66	50,752.40	50,758.49	50,757.55
Parameters	4	9	11	11	11	11

*p = <.05, ** p = < .001

performing above average on the implementation compliance index gained a mean of .62 words per minute for every point the school was above this average, while students in schools performing below the average on the index had .62 words per minute less on spring scores for each point their school was below the index.

Role of School Level Demographics in Questions 1 and 2. To determine what, if any, role school level demographics may have played in the predictiveness of the fidelity measures, an HLM analysis with the of Total Composite Score and the Implementation Compliance Index (as just-described statistically significant predictors of DIBELS oral reading fluency growth) along with four school-level demographic means was conducted separately and for each grade level. These school-level demographics included race, free/reduced lunch status, special education eligibility, and Limited English Proficiency (LEP) status. Results are summarized and presented in Table 11 for Grades 2 and Table 12 for Grade 3.

Results indicated that the Total Composite Score remained statistically significant for the DIBELS growth slope (Grade 2) and DIBELS intercept for end-of-year fluency scores (Grade 3); the coefficients associated with these effects were only slightly lowered (e.g., a coefficient change of .14 to .10 for Grade 2 and .34 to .33 for Grade 3). This indicated that this fidelity measure had very similar effects when controlling for the demographic make-up of project schools. Very similar results were found for the Implementation Compliance Index wherein the coefficients remained statistically significant and were lowered very minimally (0.18 to 0.14 in Grade 2 and 0.62 to 0.53 in Grade 3). Although demographic factors and their influence on student outcomes were

Table 11

Fixed Effects Estimates and Random Variance Estimates for Predictor Models on DIBELS Growth Grade 2 Including School-Level Demographics

Fixed effects					
Parameter	Unconditional Growth	Total Composite Index	Total Composite w/ Demographics	Implementation Compliance Index	Compliance Index with Demographics
For Intercept 2, β_{00}					
Intercept3, γ_{000}	80.23**	80.24**	80.28**	80.23**	80.27**
Composite_ME, γ_{001}		0.21	0.11		
Compliance_ME, γ_{001}				0.29	0.19
Race, ME, γ_{002}			2.79		4.30
Lunch_ME, γ_{003}			3.96		3.44
Sped_ME, γ_{004}			-70.10*		-74.31*
LEP_ME, γ_{005}			-14.08		-13.32
For Time_2 slope, π_1					
For Intercept 2, β_{10}					
Intercept 3, γ_{100}	21.31**	21.30**	21.30**	21.31**	21.30**
Composite_ME, γ_{101}		0.14**	0.10*		
Compliance_ME, γ_{001}				0.18*	0.14*
Race_ME, γ_{102}			1.04		2.76
Lunch_ME, γ_{103}			2.48		1.99
Sped_ME, γ_{104}			-26.66*		-30.77*
LEP_ME, γ_{105}			-3.56		-2.73
Random effects and model fit statistics					
Intercept, *e*	94.05	94.07	94.10	94.07	94.10
Intercept, r_0	1542.40**	1542.19**	1541.81**	1542.41**	1541.98**
Time_2 slope, r_1	50.48**	50.43**	50.37**	50.46**	50.38**
Intercept, μ_{00}	65.88**	60.28**	47.87**	62.64**	47.56**
Time_2, μ_{10}	9.63**	7.45**	5.80**	8.45**	5.97**
ICC Level 3	4.3%				
Level 3 Pseudo R^2		10.30%	28.92%	5.85%	29.11%
Deviance	62,837.19	62,827.19	62,818.38	62,832.36	62820.50
Parameters	9	11	19	11	19

*p = <.05, ** p = < .001

Table 12

Fixed Effects Estimates and Random Variance Estimates for Predictor Models on DIBELS Growth Grade 3 Including School-Level Demographics

Parameter	Unconditional Growth	Total Composite Index	Total Composite w/ Demographics	Implementation Compliance Index	Implementation Compliance Index with Demographics
Fixed effects					
For Intercept 2, β_{00}					
Intercept3, γ_{000}	99.74**	99.70**	99.75**	99.78**	99.82**
Composite_ME, γ_{001}		0.34*	0.33*		
Compliance_ME, γ_{001}				0.62*	0.53*
Race, ME, γ_{002}			-1.08		3.78
Lunch_ME, γ_{003}			12.40		10.76
Sped_ME, γ_{004}			-34.85		-48.19*
LEP_ME, γ_{005}			-26.06**		-23.62*
For Time_2 slope, π_1					
For Intercept 2, β_{10}					
Intercept 3, γ_{100}	17.89**	17.89**	17.89**	17.89**	17.89**
Composite_ME, γ_{101}		0.05	0.03		
Compliance_ME, γ_{001}				0.04	< -0.01
Race_ME, γ_{102}			3.64		4.39*
Lunch_ME, γ_{103}			6.42**		6.30**
Sped_ME, γ_{104}			-7.32		-8.40
LEP_ME, γ_{105}			-1.26		-0.95
Random effects and model fit statistics					
Intercept, *e*	80.56	80.56	80.53	80.57	80.54
Intercept, r_0	1336.07**	1336.25**	1336.14**	1336.45**	1336.37**
Time_2 slope, r_1	22.00**	22.00**	22.00**	22.00**	22.00**
Intercept, μ_{00}	68.00**	53.50**	25.45**	51.95**	24.14**
Time_2, μ_{10}	6.14**	5.78**	3.22**	6.06**	3.27**
ICC Level 3	4.90%				
Level 3 Pseudo R^2		20.04%	61.33%	21.76%	63.03%
Deviance	50,759.28	50,752.40	50,727.38	50,752.40	50,725.60
Parameters	9	11	19	11	19

*p = <.05, ** p = < .001

not a focus of this particular study, it should be noted that the pseudo R^2 for Level 3 grew substantially for both second and especially third grades with the addition of school-level demographics. In fact, adding school-level demographics to the third grade models as described above decreased the Level 3 variance by 61% and 63%, respectively.

To gather additional information on the role of school-level demographics and implementation compliance as a whole, bivariate correlations were conducted between all four measures (including the Total Composite Index) and the four school level demographic categories. Results of these correlations are found in Table 13. No statistically significant relationships were found between school level demographics and school implementation compliance scores with the exception of special education status and the professional development attendance record. These two variables were negatively correlated, $r(32) = -.385$, $p < .05$).

Table 13

Pearson Correlations Matrix of School Level Mean Demographics and Fidelity of Implementation Indices

Fidelity Index	Free/Reduced Lunch Status	Limited English Proficiency	Special Education	Race (White)
Total Composite Fidelity Score	-.085	-.032	-.244	.059
Implementation Compliance Matrix	.015	-.085	-.036	.112
Continuation Index	-.264	-.149	.115	.093
Professional Development Attendance Record	-.018	.103	-.385*	-.031

*p = <.05, ** p = < .001

Questions 1 and 2 Without School 44

As discussed earlier, one school presented itself as an outlier compared to other project schools. As noted, significant differences were found across nearly all fidelity measures including the Composite Point Index. As a result of these significant differences and a concern of undue influence, all of the HLM analyses for DIBELS outcomes across Grades 2 and 3 were rerun without School 44 to determine if different results would be obtained. Results of these analyses are presented in Table 14 for Grade 2 and Table 15 for Grade 3. Whereas the Total Composite Score was previously statistically significant in second and third grades for the DIBELS ORF growth slope and spring fluency scores, respectively, without School 44, these results were no longer statistically significant. Also of importance were differences in Level 3 variance reductions, which were less without the inclusion of School 44. As an example, the pseudo R^2 for the Total Composite model for Grade 2 was reduced from 10.30% to 4.37%, and Grade 3 from 20.04% to 7.60%. Additionally, without School 44, none of the three individual indices of implementation were fidelity statistically significant predictors of DIBELS ORF results across Grades 2 or 3. This included the Implementation Compliance Index that previously was a model that produced statistically significant results.

Role of demographics in Questions 1 and 2 Without School 44. In order to analyze the effect of school-level demographics on the model as a whole without School 44, these analyses were also rerun. Results of the analyses for Grades 2 and 3 are summarized and presented in Table 16 and Table 17, respectively. Of particular note is the even greater decrease in Level 3 variance without School 44, particularly in Grade 3.

Table 14

Fixed Effects Estimates and Random Variance Estimates for Predictor Models on DIBELS Growth Grade 2 Without School 44

Fixed effects						
Parameter	Unconditional Means	Unconditional Growth	Total Composite Index	Compliance Index	Continuation Index	Attendance Record
Intercept	58.96**	80.52**	80.53**	80.53**	80.50**	80.55**
Composite			0.21			
Compliance				0.19		
Continuation					0.67	
Attendance						-0.03
For Time_2 slope						
Intercept		21.52**	21.52**	21.52**	21.51**	21.51**
Composite			0.12			
Compliance				0.09		
Continuation					0.16	
Attendance						0.06
Random effects and model fit statistics						
Intercept, e	622.73	94.18	94.27	94.27	94.18	94.27
Intercept, r_0	1000.92**	1549.56**	1549.20**	1549.41**	1549.39**	1549.30**
Time_2 slope, r_1		51.19**	51.05**	51.06**	51.19**	51.06**
Intercept, μ_{00}	32.56*	65.03**	62.39**	64.13	60.31**	64.92**
Time_2, μ_{10}		8.51**	7.64**	8.29	8.22**	8.43**
Level 3 ICC	1.97%	4.20%				
Level 3 Pseudo R^2			4.77%	1.52%	6.81%	0.26%
Deviance	68,882.27	61,620.36	61,616.66	61,619.63	61,618.16	61,619.09
Parameters	4	9	11	11	11	11

*p = <.05, ** p = < .001

Table 15

Fixed Effects Estimates and Random Variance Estimates for Predictor Models on DIBELS Growth Grade 3 Without School 44

Parameter	Unconditional Means	Unconditional Growth	Total Composite Index	Implementation Compliance Index	Continuation Index	Attendance Record
Estimation of fixed effects						
Intercept	79.07	97.09**	97.03**	97.10**	97.08**	97.11**
Composite			0.22			
Compliance				0.43		
Continuation					0.26	
Attendance						-0.05
For Time_2 slope						
Intercept		17.76**	17.76**	17.76**	17.76**	17.76**
Composite			-0.01			
Compliance				-0.08		
Continuation					0.01	
Attendance						0.01
Random effects and model fit statistics						
Intercept, e	421.55	82.15	82.15	82.16	82.14	82.14
Intercept, r_0	1231.37**	1494.88**	1494.93**	1494.96**	1494.94**	1494.90**
Time_2 slope, r_1		25.50**	25.49**	25.47**	25.51**	25.51**
Intercept, μ_{00}	42.54**	54.90**	51.60**	49.97**	53.97**	54.76**
Time_2, μ_{10}		5.44**	5.44**	5.32**	5.44**	5.44**
Level 3 ICC	2.51%	3.63%				
Level 3 Pseudo R^2			5.47%	8.37%	2.07%	0.23%
Deviance	66,424.98	60,351.32	60,349.14	60,345.52	60,350.89	60,351.21
Parameters	4	9	11	11	11	11

*p = <.05, ** p = < .001

Table 16

Fixed Effects Estimates and Random Variance Estimates for Predictor Models on DIBELS Growth Grade 2 Including School-Level Demographics Without School 44

Fixed effects					
Parameter	Unconditional Growth	Total Composite Index	Total Composite w/ Demographics	Implementation Compliance Index	Compliance Index w/ Demographics
For Intercept 2, β_{00}					
Intercept3, γ_{000}	80.52**	80.53**	80.56**	80.53**	80.56**
Composite_ME, γ_{001}		0.21	0.06		
Compliance_ME, γ_{001}				0.19	0.12
Race, ME, γ_{002}			2.17		3.13
Lunch_ME, γ_{003}			4.02		3.65
Sped_ME, γ_{004}			-71.18*		-73.56*
LEP_ME, γ_{005}			-15.23		-14.58
For Time_2 slope, π_1					
For Intercept 2, β_{10}					
Intercept 3, γ_{100}	21.52**	21.52**	21.52**	21.52**	21.52**
Composite_ME, γ_{101}		0.12	0.07		
Compliance_ME, γ_{001}				0.09	0.08
Race_ME, γ_{102}			0.59		1.64
Lunch_ME, γ_{103}			2.54		2.25
Sped_ME, γ_{104}			-27.38*		-30.13*
LEP_ME, γ_{105}			-4.34		-3.98
Random effects and model fit statistics					
Intercept, *e*	94.18	94.27	94.28	94.27	94.21
Intercept, r_0	1549.56**	1549.20**	1548.84**	1549.41**	1549.18**
Time_2 slope, r_1	51.19**	51.05**	51.01**	51.06**	51.12**
Intercept, μ_{00}	65.03**	62.39**	49.35**	64.13**	49.03**
Time_2, μ_{10}	8.51**	7.64**	5.85**	8.29**	4.97**
ICC Level 3	4.2%				
Level 3 Pseudo R^2		4.77%	24.94%	1.52%	26.57%
Deviance	61,620.36	61,616.66	61,607.86	61,619.63	61,609.28
Parameters	9	11	19	11	19

*p = <.05, ** p = < .001

Table 17

Fixed Effects Estimates and Random Variance Estimates for Predictor Models on DIBELS Growth Grade 3 Including School-Level Demographics Without School 44

Fixed effects					
Parameter	Unconditional Growth	Total Composite Index	Total Composite w/ Demographics	Implementation Compliance Index	Compliance Index with Demographics
For Intercept 2, β_{00}					
Intercept3, γ_{000}	97.09**	97.03**	97.09**	97.10**	97.10**
Composite_ME, γ_{001}		0.22	0.07		
Compliance_ME, γ_{001}				0.43	0.21
Race, ME, γ_{002}			-3.87		-2.74
Lunch_ME, γ_{003}			11.66*		11.14*
Sped_ME, γ_{004}			-60.86*		-64.51**
LEP_ME, γ_{005}			-34.82**		-33.70**
For Time_2 slope, π_1					
For Intercept 2, β_{10}					
Intercept 3, γ_{100}	17.76**	17.76**	17.77**	17.76**	17.77**
Composite_ME, γ_{101}		-0.01	-0.04		
Compliance_ME, γ_{001}				-0.08	-0.12
Race_ME, γ_{102}			2.60		1.79
Lunch_ME, γ_{103}			6.41**		6.72**
Sped_ME, γ_{104}			-10.53		-8.53
LEP_ME, γ_{105}			-3.05		-3.75
Random effects and model fit statistics					
Intercept, *e*	82.15	82.15	82.11	82.16	82.13
Intercept, r_0	1494.88**	1494.93**	1494.14**	1494.96**	1494.24**
Time_2 slope, r_1	25.50**	25.49**	25.53**	25.47**	25.49**
Intercept, μ_{00}	54.90**	51.60**	13.25*	49.97**	12.17*
Time_2, μ_{10}	5.44**	5.44**	2.59**	5.32**	2.39**
ICC Level 3	3.63%				
Level 3 Pseudo R^2		5.47%	73.69%	8.37%	75.87%
Deviance	60,351.32	60,349.14	60,312.50	60,345.52	60,308.23
Parameters	9	11	19	11	19

*p = <.05, ** p = < .001

When race, free/reduced lunch status, special education status and LEP status were added to the Total Composite Index and Implementation Compliance Matrix models, the pseudo R^2 changed from 5.47% to 73.69% and 8.37% to 75.87%, respectively, when compared to the unconditional growth model.

Questions 3 and 4 With Full Sample

The second series of models were targeted at answering research questions 3 and 4, which are reproduced below:

3. To what extent is school-level variance on the Stanford Achievement Test-10 (second grade) and Oregon Test of Knowledge (third grade) in Oregon Reading First schools associated with higher levels of implementation of the Schoolwide Reading Model as measured by a composite index of three indices of implementation compliance?
4. To what extent do the three different indices of implementation compliance independently explain student performance on summative, end-of-year reading achievement measures of the Stanford Achievement Test-10 and Oregon Test of Knowledge and Skills?

In order to answer Questions 3 and 4, a two-level HLM model was utilized with students nested within 34 schools. Question 3 used the combined index of implementation compliance as the Level 2 predictor of *SAT-10* and the *OAKS Reading/Literature* outcomes. Within Question 4, each of the three indices of implementation compliance separately served as Level 2 predictors of outcomes on the *SAT-10* and the *OAKS Reading/Literature* assessments.

Grade 2. Initially, an HLM unconditional means model was conducted, this time to provide information about how much variation in SAT-10 outcomes existed between the two levels of the model as well as test the hypothesis that the variability was null. The Intraclass Correlation Coefficient (ICC), as shown in Table 18, indicated that about

Table 18

Fixed Effects Estimates and Random Variance Estimates for Models Predicting Grade 2 SAT-10 Results

Fixed Effects					
Parameter	Unconditional Means	Composite Index	Compliance Index	Continuation Index	Attendance Record
For Intercept 1					
Intercept 3	584.25**	584.23**	584.25**	584.22**	584.27**
Composite		0.11			
Compliance			0.21		
Continuation				0.84	
Attendance					-0.16
Random effects and model fit statistics					
Intercept, r	1752.38	1752.51	1752.53	1752.17	1752.40
Intercept, μ_0	103.50**	105.52**	105.21**	100.09**	106.55**
Level 2 ICC	5.58%				
Level 2 Pseudo R^2		-1.95%	-1.65%	3.29%	-2.95%
Deviance	21,185.45	21,184.79	21,183.60	21,180.78	21,183.77
Parameters	2	2	2	2	2

*p = <.05, ** p = < .001

94.42% of the variance in DIBELS growth was between students (Level 1) and 5.58% of the variance was between schools (Level 2). Once again, although small, the Level 2 variance was statistically significant.

As Table 18 also indicates, unlike reading fluency scores, the Total Composite Point fidelity index was not a statistically significant predictor of SAT-10 outcomes at Grade 2. The addition of this predictor slightly increased the variance at Level 2.

Similarly, when the Implementation Compliance Index, Continuation Index and Professional Development Attendance Record Index were added to the model independently, none of the three measures were statistically significant predictors of SAT-10 outcomes. In addition, little to no reduction in variance occurred with these predictors in the model. Overall, a good model fit was not indicated.

Grade 3. Similar analyses were conducted using Grade 3 OAKS-Reading data in order to answer Questions 3 and 4. Table 19 highlights these results. The null model revealed an Intraclass Correlation Coefficient (ICC) of .08. Thus, approximately 8% of the variance in OAKS-Reading scores was between schools, and about 92% of the variance in scores was at the student level. Because significant variance existed at both levels of the data structure, HLM was again a suitable approach for modeling the data. As noted in Table 19, the Composite Point Total fidelity index was not a statistically significant predictor of OAKS-Reading outcomes at Grade 3. The addition of this predictor slightly increased rather than decreased the variance at Level 2. The Implementation Compliance Index, Continuation Index and Professional Development Attendance Record Index were then added to the model independently to estimate the Level 2 variance explained by each separate model. Similar to the results using the Grade 2 SAT-10, none of the three models produced statistically significant results. In addition, little to no reduction in variance at Level 2 occurred as a result of adding these predictors in the model.

Role of Demographics in Questions 3 and 4. To determine what, if any, role school level demographics may have played in the predictability of fidelity measures on SAT-10 and OAKS-Reading outcomes of fidelity measures, an HLM analysis using the

Table 19

Fixed Effects Estimates and Random Variance Estimates for Models Predicting Grade 3 OAKS-Reading Results

Estimation of fixed effects					
Parameter	Unconditional Means	Composite Index	Compliance Index	Continuation Index	Attendance Record
Intercept	209.41*	209.41**	209.41**	209.41**	209.40**
Composite		0.07			
Compliance			0.05		
Continuation				0.05	
Attendance					0.12
Random effects and model fit statistics					
Intercept, r	103.65	103.66	103.66	103.65	103.66
Intercept, μ_0	9.04**	8.74**	9.24**	9.35**	8.73**
Level 2 ICC	8.02%				
Level 2 Pseudo R^2		3.32%	-2.21%	-3.42%	3.42%
Deviance	15,254.91	15,259.06	15,259.34	15,258.35	15,257.82
Parameters	2	2	2	2	2

*p = <.05, ** p = < .001

Total Composite Index together with four Level 2 school-level demographic means, including race, free/reduced lunch status, special education eligibility, and LEP status was conducted. Table 20 summarizes these results. The addition of school-level demographics once again lowered Level 2 variance to a noticeable degree in both grade levels. Approximately 36% to 26% of the variance was accounted for at Level 2 for Grades 2 and 3, respectively, with the addition of school level demographics when compared to the null model.

Questions 3 and 4 Without School 44

All of the above analyses with Questions 3 and 4 were rerun without School 44. Findings were very similar to those HLM analyses performed with the SAT-10 and OAKS-Reading at both grade levels with all schools. No statistically significant findings or significant differences in Level 2 variance were found between all schools and the

models rerun without School 44 as is illustrated with the full results in Table 21 for Grade 2 and Table 22 for Grade 3.

Table 20

Fixed Effects Estimates and Random Variance Estimates for Predictor Models of SAT-10 and OAKS-Reading Results Including School-Level Demographics

	Fixed effects					
	Grade 2			Grade 3		
	Unconditional Growth	Total Composite Index	Total Composite w/ Demographics	Unconditional Growth	Total Composite Index	Total Composite w/ Demographics
For Intercept 2, β_0						
Intercept3, γ_{00}	584.25**	584.23**	584.41**	209.41*	209.41**	209.42
Composite_ME, γ_{01}		0.11	0.02		0.07	0.01
Race, ME, γ_{02}			-1.17			4.82*
Lunch_ME, γ_{03}			0.02			1.67
Sped_ME, γ_{04}			-72.28*			-15.86
LEP_ME, γ_{05}			-39.96**			-5.13*
	Random effects and model fit statistics					
Intercept, level-1, *r*	1752.38	1752.51	1752.85	103.65	103.66	103.66
Intercept, u_0	103.50**	105.52**	65.79**	9.04**	8.74**	6.67**
ICC Level 2	5.58%			8.02%		
Level 2 Pseudo R^2		-1.95%	36.43%		3.32%	26.22%
Deviance	21,185.45	21,184.79	21,142.02	15,254.91	15,259.06	15,229.93
Parameters	2	2	2	2	2	2

*p = <.05, ** p = < .001

Table 21

Fixed Effects Estimates and Random Variance Estimates for Models Predicting Grade 2 SAT-10 Results Without School 44

Fixed Effects					
Parameter	Unconditional Means	Composite Index	Compliance Index	Continuation Index	Attendance Record
For Intercept 1					
Intercept 3	584.19**	584.17**	584.21**	584.16**	584.20**
Composite		0.25			
Compliance			0.40		
Continuation				0.86	
Attendance					-0.18
Random effects and model fit statistics					
Intercept, *r*	1760.67	1760.78	1760.80	1760.44	1760.68
Intercept, μ_0	106.95**	106.55**	106.74 **	103.31**	110.34**
Level 2 ICC	5.73%				
Level 2 Pseudo R^2		0.37%	0.20%	3.40%	-3.17%
Deviance	20,.788.13	20,780.21	20,779.30	20,777.39	20,780.10
Parameters	2	2	2	2	2

*p = <.05, ** p = < .001

Table 22

Fixed Effects Estimates and Random Variance Estimates for Models Predicting Grade 3 OAKS-Reading Results Without School 44

Estimation of fixed effects					
Parameter	Unconditional Means	Composite Index	Compliance Index	Continuation Index	Attendance Record
Intercept	209.47*	209.46**	209.47**	209.47**	209.46**
Composite		0.09			
Compliance			0.03		
Continuation				0.04	
Attendance					0.15
Random effects and model fit statistics					
Intercept, *r*	104.30	104.31	104.31	104.30	104.31
Intercept, μ_0	9.24**	8.99**	9.57**	9.58**	9.05**
Level 2 ICC	8.14%				
Level 2 Pseudo R^2		2.71%	-0.04%	-0.04%	2.06%
Deviance	14,787.08	14,790.72	14,791.25	14,790.53	14,789.97
Parameters	2	2	2	2	2

*p = <.05, ** p = < .001

CHAPTER V

DISCUSSION

The purpose of this study was to make use of indicators of levels of implementation collected during the enactment of the Oregon Reading First program in order to examine whether variation of implementation of the components of the Schoolwide Reading Program predicted better outcomes for students and schools. In particular, the aim of this study was to determine the extent to which each of three different types of measures of implementation compliance, as well as a combined index of these measures collected during the implementation of Oregon Reading First, explained school-level variance in student improvement across various measures of reading skills.

Summary of Results and Implications

As stated by Odom and Cook (2013), the potential benefit of evidence-based practices is bound by the quality, reach, and maintenance of implementation. While this certainly is a well-grounded argument, mixed results, at best, were found in this particular study that both align and diverge from previous studies linking fidelity measures with student outcomes. A discussion of these results follows.

Question 1 – Composite Measure of DIBELS Growth

In both second and third grades, the Composite Point Index, a combined score of the three different measures of fidelity, was a statistically significant predictor of oral reading fluency growth. Although significant, the associations at both grade levels were quite small for the Grade 2 slope (an average 1.57 words per minute additional growth for a student in a school performing one standard deviation above the mean on this index)

and the Grade 3 intercept (an average 3.80 words per minute additional end-of-year growth for a third grade student in a school performing one standard deviation above the mean on the composite index), respectively. The composite index explained approximately 10% to 20% additional variance between schools in these two grade levels compared to the null growth models that, to begin with, had small ICC's of 0.43 and 0.49. These results indicate factors other than implementation fidelity may explain variance that existed between schools.

Question 2 – Single Measure Predictors of DIBELS Growth

Results indicate that, of the three single measures of fidelity, only the Implementation Compliance Index was a statistically significant predictor of Oral Reading Fluency. Table 2 presented earlier lists components of the Implementation Compliance Index, which, in essence, is an adherence measure and was verified through required deliverables and classroom observations. As was the case for the Total Composite Index, a statistically significant effect was found for the DIBELS growth slope for Grade 2, and the effect was found on the end-of-year spring DIBELS score for Grade 3.

The magnitude of the statistically-significant coefficients for the Implementation Compliance Index was, again, relatively small. For example, for students in Grade 2, a school one standard deviation above average for this fidelity measure would result in additional approximate 1.3 words per minute additional growth on the slope; an additional 3.9 words per minute gain would ensue for third grade students' spring DIBELS scores. Although this additional growth is relatively small, in schools with

many students who struggle to read, these effects may or may not have practical significance.

The Peculiar Case of School 44

The small, but statistically significant results for Questions 1 and 2 are tempered by the fact that the absence of outlier School 44 changed the results of the HLM analyses when rerun. Across both second and third grades, without School 44, those models that previously did produce statistically significant outcomes were no longer statistically significant.

One possibility for this result is the previously-cited phenomenon of implementation threshold effects (Durlak & DuPre, 2008). That is, once a certain level of fidelity is obtained, higher levels may not always lead to significantly better outcomes, particularly if the intervention's core components have been effectively delivered. Durlak and DuPre suggested that implementation fidelity that falls somewhere within the 60% to 80% range may be acceptable if the core components of the program are identified and put in place by those responsible for implementation. It is noteworthy that out of the 34 project schools, 32 schools fell above this threshold range when examining total percent of implementation as measured by the Total Composite Index. Another school fell slightly below this range. Percent of implementation ranged from 55% to 94% with these 33 schools. The overall percent of implementation for School 44 was 27%. Similar results occurred with the other previously-cited statistically significant result using the Implementation Compliance Index. Once again, 32 out of the 34 project schools fell within or above the 60% to 80% fidelity range using this measure. School 44's overall percent of implementation on this measure was 20%. It should be noted that

core components of the Schoolwide Reading Model were clearly identified by Oregon Reading First staff, and schools were given extensive support through ongoing on-site professional development and technical assistance.

Another hypothesis for the small statistically significant predictiveness, as well as lack of statistically significant effects without School 44, is the extensive on-site professional development and technical assistance received by Oregon Reading First school personnel from Oregon Reading First center staff during implementation of the Schoolwide Reading Model. Basaraba (2011), in a separate study using Oregon Reading First data, discusses this considerable support. Schools received ongoing professional development throughout the first two years of implementation on each of the seven components of the Schoolwide Reading Model as identified earlier. Schools also received ongoing technical assistance from State Reading First personnel assigned to specific schools for ongoing support. Additionally, school coaches were hired at each participating Reading First school to provide internal support to teachers to implement the model. These coaches were trained by Oregon Reading First state personnel.

Implementation data indicates this wide-reaching support resulted in a large majority of Oregon Reading First schools implementing elements of the Schoolwide Reading Model with a high rate of fidelity. In reality, fidelity of implementation to the various components of a multi-tiered system of supports such as the Schoolwide Reading Model may look very different in schools without access to this type of far-reaching assistance. As a result, the predictiveness of these fidelity measures on student reading outcomes may look quite different with more variability in fidelity of implementation to the model.

Finally, another potential explanation for the limited results with and without School 44 is that the measures themselves were not built to capture day-to-day implementation of the Schoolwide Reading Model and those instructional strategies and nuances that may produce significant differences in outcomes for students. This is briefly discussed in the Limitations section of this chapter. As stated earlier, all of the measures used to examine fidelity were, for all practical purposes, fidelity compliance or adherence-based rather than process or quality-based measures. For example, the Professional Development Attendance Record captured attendance and learning at scheduled events. However, this measurement tool did not capture if the new learnings were implemented on a day-to-day basis. The Continuation Application was a self-report, and weaknesses in self-reports were discussed earlier by researchers such as Century et al. (2010) and Gresham et al. (2000). This measure was also a reflection of how well the school complied with implementing the Schoolwide Reading Model.

Additionally, although the third measure, the Implementation Compliance Matrix, did incorporate a number of different artifacts and observations relevant to implementation of Oregon Reading First, a weakness in this measure was that for the most part, fidelity points were assigned for mere compliance and the delivery of required artifacts, and not for quality of those fidelity artifacts. Although these measures may well have been necessary to capture overall implementation of the various required components of the program, they may not have been sufficient enough to gather the type of information needed to differentiate those specific components of implementation that related more directly to improved student outcomes.

Questions 3 and 4 –Total Composite Index and Single Model Predictors of SAT-10 and OAKS-Reading Outcomes

Analyses of the relationship between all three single fidelity measures as well as the total composite scores with overall reading outcomes as measured by the SAT-10 in Grade 2 and the OAKS-Reading/Literature in Grade 3 produced no statistically significant results. Additionally, variance at Level 2 actually increased in three out of the four models in Grade 2 and two out of the four models in Grade 3. Results were the same for analyses conducted with and without School 44. In addition to those reasons discussed previously, several other possibilities exist as to why these results were obtained.

One possibility for the lack of predictiveness, particularly with the SAT-10 and OAKS-Reading outcomes, is that instructional activities may have been focused more on the development of access skills, with less rigor devoted to comprehension. Thus, reading comprehension outcomes in schools implementing Reading First with strong fidelity may not have differentiated themselves from schools implementing with less fidelity. This possibility aligns with similar theories proposed by researchers that have studied results from other state Reading First results. For example, in a study of reading comprehension results from Michigan's implementation of Reading First (Carlisle, Cortina, & Zeng, 2010), researchers theorized that instruction in the five components of reading required by Reading First was "not sufficiently infused with cognitively challenging instruction of the kind that is thought to contribute to academic achievement" (p.66). In a study on Reading First outcomes in the state of Florida (Connor, Jakobsons, Crowe, & Meadows, 2009), results indicated that overall, children in Reading First

classrooms were achieving grade expectations in oral reading fluency, and most first graders demonstrated expected reading comprehension skills by the end of the school year. However, second and third graders did not experience the same results with comprehension. One of the theories advanced by the authors was that the instructional strategies used may not have been explicitly focused on instruction and practice in comprehension strategies.

Another theory for the lack of statistically significant results using fidelity measures and the SAT-10 and OAKS-Reading assessments is that improving reading comprehension skills is simply more difficult and takes more time than improving lower-level skills. Given the complex nature of reading comprehension, it may in fact take more time for students to learn, practice and implement reading comprehension strategies independently than the measurement schedule allowed in order to capture these effects. In other words, it may be that not enough lag time occurred between initial instruction on reading comprehension skills and measuring implementation of those skills. The importance of allowing time for reading comprehension instruction to take hold was emphasized by Berkeley, Scruggs and Mastropieri (2010) in a meta-analysis of reading comprehension strategies for students with learning disabilities. They suggested future research should be conducted to provide more evidence on the effects of longer term implementation of reading comprehension instruction on norm-referenced measures of reading.

Role of Demographics in the Relationship Between Fidelity and Student Outcomes

In both second and third grades, the coefficients for the statistically significant models using the Total Composite Score and Implementation Compliance Index were

impacted very little when accounting for school level demographics. This indicates the small but statistically significant effects that were found remained so even after accounting for demographics. However, and perhaps more importantly, the decrease in Level 3 variance when adding school-level demographics to these models and substantial increases in pseudo R^2 calculations suggests that demographics may have played an important role in differences in outcomes between schools outside of the predictiveness of implementation fidelity. Even with extensive support provided to schools in this project on developing of a multi-tiered framework of support and using evidence-based instructional practices, it appears that school-level demographics most likely still contributed to the differences in school outcomes.

If this were the case, it would align with prior research conducted on outcomes for students in schools within the Michigan Reading First program referenced earlier (Carlisle, Cortina, & Zeng, 2010). These authors noted that smaller percentages of free and reduced lunch status students performed at or above grade level compared to peers. They also noted that the effects of poverty were enduring even when these schools were given additional support. The authors also found lower performance levels for students with disabilities, and that the performance gap did not narrow significantly over time. Finally, their results for LEP students showed that the percentage of LEP students reading at or above grade level was similar to that of non-LEP students in first grade, but that the gap widened somewhat between LEP and non-LEP students in second and third grades.

Limitations

A number of limitations constrain generalization of the findings of this study as well as moderate interpretation of the findings. To begin, the extant data that was utilized

as measures of fidelity in this study was not collected specifically for the purpose of measuring fidelity of implementation. Rather, the data was originally collected to aid in making decisions related to continuation of funding for Oregon Reading First schools as well as inform Oregon Reading First staff as to the type of support that was needed by each school. Relatedly, the data used for measures of fidelity within this study was collected well over 10 years ago. The types of data collected do parallel current research on general ways to measure overall implementation compliance, such as permanent products, observations and checklists. However, if collecting data on implementation fidelity at the present time, specifically for implementation of multi-tiered system of support, more and most likely different measures might be used to more clearly align to specific theories of implementation such as those discussed by Fixen et al. (2005), Sullivan, Blevins, & Kauth (2008) and Ogden and Fixen (2014), as well as capture more subtle qualitative gradations of implementation beyond the basic framework level that relate specifically to improved outcomes. Important components of schoolwide literacy systems such as a culture change related to the use of data to inform instruction, intensity of instruction for struggling students, and/or active principal leadership related to instruction are examples of the types of key ingredients that could be integrated into implementation fidelity indices.

A second limitation of this study relates to generalizability of findings due to missing data and missing data patterns. As previously mentioned, the SPSS Missing Values Analysis found that data was not missing at random at both grade levels. An analysis of missing data patterns through lead to the conclusion that students with the pattern of scores on all four measures (XXXX) were statistically significantly different

and higher performing than students with almost all other missing patterns across various DIBELS time point measurements and both the SAT-10 and OAKS-Reading measures. In addition, nearly 23% of the data was missing for both the SAT-10 and OAKS-Reading. Together this indicates that students within the HLM analyses for overall reading competency may well have been the higher-performing students within the study. As a result, findings from this study are biased, and generalization of findings are limited to schools within this study.

Another important concern for generalizability of results and thus a limitation of this particular study is the extensive on-site professional development and technical assistance received by Oregon Reading First schools from Oregon Reading First center staff during implementation of the Schoolwide Reading Model, as previously discussed. Other schools attempting to implement multi-tiered systems of support may experience very different results depending upon the amount and type of assistance received from outside experts. In many cases, this support will most likely not be nearly as intensive and extensive as schools participating in Oregon Reading First. As a result, this limits the generalizability of results of this particular study to other schools implementing an MTSS systems.

Finally, another obvious limitation of this study is the fact that it did not use an experimental design. Within Reading First national guidelines, this was simply not possible. As noted earlier, random assignment remains the most reliable technique for justifying causal inference. It provides the logically most valid and efficient causal counterfactual. Consequently, results are more credible than those from other quasi- or non-experimental methods (Steiner, Wroblewski, & Cook, 2009). U. S. Department of

Education policy clearly defined the specific criteria that were used to select schools that were part of the overall Reading First program and thus created a limited set of schools for which generalization of results could potentially apply to. The fact that these schools were not randomly selected, could volunteer to opt out of the program if they so desired and potentially had somewhat similar characteristics could also be a factor as to why implementation fidelity did not play a larger role in predicting student outcomes. The results may have been much different with randomly-selected schools. From a larger perspective, the fact that policy initiatives often include selection criteria, contingent funding, and specific participation requirements may make these initiatives a poor context for research on the effects of implementation fidelity in general.

Conclusion and Implications

The goals of this study were to determine the relationship between implementation compliance measures used during Reading First and outcome measures for students within each of these Reading First schools in Oregon, as well as add to the literature base in better understanding fidelity of implementation as it relates to student reading outcomes within multi-tiered systems of support. Ultimately, limited relationships, at best, were found in analyzing this association, although several findings from this study can add to the overall understanding of measuring fidelity of implementation.

First, findings from the fluency component of this study add some support to the theory of implementation threshold effects. The difference in HLM analyses with and without School 44, the outlier school, supports the premise that once a certain level of implementation is attained, higher levels may not always lead to significantly better

outcomes. What appears to be most important is identifying the most critical components of an initiative or program and ensuring these components are effectively implemented. This also implies that, above and beyond implementation of the key essential components of a practice or program, some flexibility in implementation based upon specific circumstances, such as local context or needed cultural adaptations, would not necessarily adversely affect student outcomes.

Next, and again not part of the original focus of this study, findings from this investigation emphasize the continued need to support research in early literacy that lessens the effects of poverty, limited English proficiency and disabilities on literacy outcomes, particularly reading comprehension. The large pseudo R^2 percentages that were obtained when demographics were added to various models indicate that a significant amount of variance between schools was explained by demographics despite the use of evidence-based practices and strong support for implementation of these practices. This issue most likely is related more to the development, selection and use of evidence-based practices that more intensively address the unique needs of students that fall into these disaggregated subcategories, rather than implementation fidelity concerns. Continued research that addresses the unique instructional needs of these students, particularly students with disabilities and students with limited English proficiency, will be critical to improve the effectiveness of MTSS systems.

Finally, continued development and refinement of measures of implementation fidelity appears important. Measures that encompass both adherence and those qualitative components of MTSS systems that make important differences for reading improvement seem particularly important. Additionally, aligning measures of

implementation with evidence-based practices that specifically focus on the reading improvement for schools with high concentrations of at-risk students also appears to be an important need.

It is clear additional research is needed to better understand the relationship between implementation fidelity and student outcomes, particularly as it relates to multi-tiered systems of support. Future research may best be undertaken without some of the limitations discussed earlier, such as conducting randomized control studies without the incentive of continued funding and outside the context of policy and mandates. In the field of education, the use of multi-tiered systems of support within schools will only continue to grow. It is, therefore, critical that research on the effectiveness of implementation of these systems as they relate to student outcomes continues to expand as well.

APPENDIX A

OREGON READING FIRST CONTINUATION APPLICATION:

COHORT A SCHOOLS

Section A: Summary and Analysis of Student Performance

1. Using your DIBELS data from last year (2003-2004), identify and document one essential instructional component (e.g., phonemic awareness, alphabetic principle, oral reading fluency) in which all or most of your students met the DIBELS benchmark goal. For example: (a) your 2004 winter to spring Summary of Effectiveness DIBELS reports indicate that all first grade strategic students reached the end ORF goal of 40 wcpm; or (b) your DIBELS 2004 winter to spring Summary of Effectiveness reports indicate that 80% of all kindergarten students reached the end of year PSF goal of 35. Please attach all data report summaries used in this analysis.

2. Using your end of year DIBELS data from last year (2003-2004), identify and document one area (e.g., phonemic awareness, alphabetic principle, oral reading fluency) in which your students did not meet the DIBELS benchmark goal and further improvement is most strongly needed. For example: (a) your 2004 winter to spring Summary of Effectiveness DIBELS reports indicate that only 18% of second grade strategic students reached the end of year ORF goal of 90 wcpm; or (b) your 2004 end of year School Report indicates that 50% of third grade students are at risk for reading difficulty. Please attach all data report summaries used in this analysis.

3. Explain how you addressed the needed improvement in item #2 in the fall of 2004. For example: (a) all second grade strategic students received additional instruction in phonics using Touchphonics and additional instruction in fluency using Read Naturally; or (b) all third grade intensive students received intervention daily in Corrective Reading for 45 minutes. What changes in your DIBELS data have you seen so far this year in the area needing further improvement? Please attach all data reports and fall CSI maps used in this analysis. Please highlight the changes made to the CSI maps based on your data analysis.

Section B: Fidelity of Implementation

1. Explain how often and how long grade level teams meet to analyze student data. Describe what student data are discussed and how instructional adjustments are made based on data? How often does the principal participate in these meetings? Please attach a copy of your 2004-2005 meeting schedule and meeting agendas for October and November.

2. Is the reading coach expected to perform duties that fall outside of the Oregon Reading First coach's job description? If yes, please describe those duties and the frequency with which they are performed. Attach a copy of the coach's schedule.

3. The fall fidelity observations were to be submitted to the regional coordinator for all teachers no later than January 15. How does the coach use the information collected on the fidelity observations? Please verify that all fidelity observations were completed and submitted by the due date. If they were not, please explain.

4. One of the requirements for Oregon Reading First is to collect lesson progress reports (LPRs) on a monthly basis. Please describe your process for collecting and using LPRs. Attach a copy of your second grade November 2004 lesson progress report.

Section C: Leadership

1. How often does the school principal observe instruction in the classroom during Reading First time? What procedures are used by the principal to determine which classrooms to observe? On average, how long do these observations last? Does the principal use specific observation forms or instruments? If yes, please describe these procedures and attach a blank copy of the observation forms.

2. One of the requirements of Reading First is that ***all*** Reading First staff, including principals, is to attend all Reading First Institutes (IBRs) on Beginning Reading and all Leadership Sessions. Have you have satisfied this requirement? If not, please explain.

Section D: District Support

1. How often does the district team meet for the purpose of analyzing district wide Reading First data? How does the district determine if schools are implementing Reading First as intended? What steps are taken to assist schools that are not on track to meet end of year reading performance goals?

2. One of the requirements of Reading First is that ***all*** Reading First district team members attend all Reading First Institutes on Beginning Reading and all Leadership Sessions. Have you satisfied this requirement? If not, please explain.

3. Explain the district's established plan for ongoing communication and collaboration with school principals and reading coaches to maintain a shared focus on Reading First.

Section E: Budget

Approval of Mid-Term Budget Report (submitted to ODE by January 21, 2005)

APPENDIX B

ADDITIONAL TABLES

Table B1

Missing Data by Count and Expectation Grade 2 SY 2004-2005

	DIBELS Beginning ORF					DIBELS Middle ORF					DIBELS Ending ORF					SAT-10				
ID	Count		Expected		Diff	Count		Expected		Diff	Count		Expected		Diff	Count		Expected		Diff
	N	%	*N*	%	%	*N*	%	*N*	%	%	*N*	%	*N*	%	%	*N*	%	*N*	%	%
2	7	6.0	12.3	10.6	4.6	14	12.1	11.4	9.8	-2.3	19	16.4	13.6	11.7	-4.7	40	34.5	26.3	22.7	-11.8
3	8	10.7	7.9	10.5	0.2	1	1.3	7.4	9.9	8.6	9	12.0	8.8	11.7	-0.3	12	16.0	17.0	22.7	6.7
6	6	9.8	6.5	10.7	-0.9	6	9.8	6.0	9.8	0.0	6	9.8	7.2	11.8	2.0	8	13.1	13.8	22.6	9.5
7	3	5.3	6.0	10.5	5.2	2	3.5	5.6	9.9	6.4	5	8.8	6.7	11.8	3.0	6	10.5	12.9	22.6	12.1
8	7	7.6	9.7	10.5	2.9	10	10.9	9.1	9.9	-1.0	9	9.8	10.8	11.7	1.9	16	17.4	20.9	22.7	2.7
9	10	18.2	5.8	10.5	-7.7	3	5.5	5.4	9.8	4.3	7	12.7	6.5	11.8	-0.9	11	20.0	12.5	22.7	2.7
10	6	7.1	9.0	10.6	3.5	9	10.6	8.4	9.8	0.8	10	11.8	10.0	11.8	0.0	18	21.2	19.3	22.7	1.5
11	6	8.1	7.8	10.5	2.4	5	6.8	7.3	9.9	3.1	7	9.5	8.7	11.8	2.3	16	21.6	16.8	22.7	1.1
14	2	2.8	7.5	10.6	7.8	0	0.0	7.3	9.9	9.9	0	0.0	8.3	11.7	11.7	7	9.9	16.1	22.7	12.8
16	4	12.1	3.5	10.6	-1.5	5	15.2	3.2	9.7	5.5	5	15.2	3.9	11.8	-3.4	6	18.2	7.5	22.7	4.5
20	8	8.9	9.5	10.6	1.7	10	11.1	8.9	9.9	-1.2	17	18.9	10.6	11.8	-7.1	20	22.2	20.4	22.7	0.5
21	2	3.9	5.4	10.6	6.7	3	5.9	5	9.8	3.9	5	9.8	6.0	11.8	2.0	8	15.7	11.6	22.7	7.0
22	6	8.2	7.7	10.5	3.5	10	13.7	7.2	9.9	-3.8	12	16.4	8.6	11.8	-4.6	18	24.7	16.6	22.7	-2.0

25	15	18.8	8.5	10.6	-8.2	4	5.0	7.9	9.9	4.5	8	10.0	9.4	11.8	1.8	16	20.0	18.2	22.8	2.8
27	11	12.1	9.6	10.5	-1.6	15	16.5	9.0	9.9	-6.7	16	17.6	10.7	11.8	-5.8	21	23.1	20.6	22.6	-0.5
29	6	8.8	7.2	10.6	1.8	3	4.4	6.7	9.9	5.5	6	8.8	8	11.8	3.0	18	26.5	15.4	22.6	-3.9
30	4	4.8	8.8	10.6	5.8	5	6.0	8.2	9.9	3.9	7	8.4	9.8	11.8	3.4	16	19.3	18.8	22.7	3.4
34	3	6.8	4.7	10.7	3.9	3	6.8	4.3	9.8	3.0	3	6.8	5.2	11.8	5.0	15	34.1	10.0	22.7	-11.4
38	10	14.3	7.4	10.6	-3.7	8	11.4	6.9	9.9	-1.5	8	11.4	8.2	11.7	0.3	16	22.9	15.9	22.7	-0.2
41	8	11.8	7.2	10.6	-1.2	9	13.2	6.7	9.7	-3.5	3	4.4	8.0	11.8	7.4	16	23.5	15.4	22.6	-0.9
43	7	8.6	8.6	10.6	2.0	9	11.1	8.0	9.9	-1.2	16	19.8	9.5	11.7	-8.1	29	35.8	18.4	22.7	-13.1
44	7	13.2	5.6	10.6	-2.6	6	11.3	5.2	9.8	-1.5	6	11.3	6.2	11.7	0.4	13	24.5	12.0	22.6	-1.9
47	13	15.1	9.1	10.7	-4.4	14	16.3	8.5	9.8	-6.5	5	5.8	10.1	11.7	5.9	12	14.0	19.5	22.7	8.7
48	17	17.0	10.6	10.6	-6.4	18	18.0	9.8	9.8	-8.2	16	16.0	11.8	11.8	-4.2	26	26.0	22.7	22.7	-3.3
49	23	21.5	11.3	10.6	-10.9	14	13.1	10.5	9.8	-3.3	8	7.5	12.6	11.8	4.3	23	21.5	24.3	22.7	1.2
50	7	11.5	6.5	10.7	-.8	8	13.1	6.0	9.8	-3.3	8	13.1	7.2	11.8	-1.3	12	19.7	13.8	22.6	2.9
51	10	9.5	11.1	10.6	1.1	11	10.5	10.3	9.8	-0.7	13	12.4	12.3	11.7	-0.7	21	20.0	23.8	22.7	2.7
55	2	1.7	12.3	10.6	8.9	6	5.2	11.4	9.8	4.6	8	6.9	13.6	11.7	4.8	19	16.4	26.3	22.7	6.3
57	9	15.5	6.1	10.5	-10.0	5	8.6	5.7	9.8	1.2	8	13.8	6.8	11.7	-2.1	19	32.8	13.2	22.8	-10.0
58	14	16.5	9.0	10.6	-5.9	8	9.4	8.4	9.8	0.4	7	8.2	10.0	11.8	3.6	24	28.2	19.3	22.7	-5.5
60	14	14.3	10.4	10.6	-3.7	9	9.2	9.6	9.8	0.6	12	12.2	11.5	11.7	-0.5	26	26.5	22.2	22.7	-3.8
62	13	10.5	13.1	10.6	0.1	9	7.3	12.2	9.8	2.5	18	14.5	14.6	11.8	-2.7	37	29.8	28.1	22.7	-7.1
66	7	8.0	9.3	10.6	2.6	15	17.0	8.7	9.9	-7.1	16	18.2	10.3	11.7	-6.5	24	27.3	20.0	22.7	-4.6
68	6	11.1	5.7	10.6	-0.5	4	7.4	5.3	9.8	2.4	9	16.7	6.4	11.9	-4.8	13	24.1	12.3	22.8	-1.3

Table B2

Missing Data by Count and Expectation Grade 3 SY 2004-2005

	DIBELS Beginning ORF					DIBELS Middle ORF					DIBELS Ending ORF					SAT-10				
ID	Count		Expected		Diff	Count		Expected		Diff	Count		Expected		Diff	Count		Expected		Diff
	N	%	*N*	%	%	*N*	%	*N*	%	%	*N*	%	*N*	%	%	*N*	%	*N*	%	%
2	10	10.5	9.5	10.0	-0.5	6	6.3	8.8	9.2	2.9	11	11.6	11.4	12.0	0.4	18	18.9	21.6	22.7	3.9
3	8	10.5	7.6	10.0	-0.5	7	9.2	7.0	9.2	0.0	3	3.9	9.1	12.0	8.1	7	9.2	17.3	22.8	10.4
6	8	12.7	6.3	10.0	-2.7	3	4.8	5.8	9.2	4.4	5	7.9	7.6	12.1	4.2	7	11.1	14.3	22.7	11.6
7	5	7.0	7.1	10.0	3.0	4	5.6	6.6	9.3	3.7	8	11.3	8.5	12.0	0.8	18	25.4	16.1	22.7	-2.7
8	6	7.1	8.5	10.0	2.9	4	4.7	7.9	9.3	4.6	8	9.4	10.2	12.0	2.6	21	24.7	19.3	22.7	-2.0
9	8	11.3	7.1	10.0	-1.3	8	11.3	6.6	9.2	-2.1	10	14.1	8.5	12.0	-2.1	26	36.6	16.1	22.7	-13.9
10	9	9.9	9.1	10.0	0.1	19	20.9	8.4	9.2	-11.7	14	15.4	10.9	12.0	-3.4	29	31.9	20.7	22.7	-9.2
11	10	12.3	8.1	10.0	-2.3	10	12.3	7.5	9.3	-3.0	10	12.3	9.7	12.0	-0.3	24	29.6	18.4	22.7	-6.9
14	5	7.6	6.6	10.0	2.4	6	9.1	6.1	9.2	0.1	5	7.6	7.9	12.0	4.4	8	12.1	15.0	22.7	10.6
16	7	13.5	5.2	10.0	-3.5	5	9.6	4.8	9.2	-0.4	11	21.2	6.2	12.0	-9.2	14	26.9	11.8	22.7	-4.2
20	6	6.9	8.7	10.0	3.1	8	9.2	8.0	9.2	0.0	15	17.2	10.5	12.1	-5.1	38	43.7	19.8	22.8	-20.9
21	11	14.5	7.6	10.0	-4.5	9	11.8	7.0	9.2	-2.6	13	17.1	9.1	12.0	-5.1	22	28.9	17.3	22.8	-6.1
22	7	9.3	7.5	10.0	0.7	3	4.0	6.9	9.2	5.2	7	9.3	9.0	12.0	2.7	14	18.7	17.1	22.8	4.1
25	14	18.2	7.7	10.0	-8.2	3	3.9	7.1	9.2	5.3	16	20.8	9.3	12.1	-8.7	17	22.1	17.5	22.7	0.6
27	8	9.9	8.1	10.0	0.1	11	13.6	7.5	9.3	-4.3	14	17.3	9.7	12.0	-5.3	17	21.0	18.4	22.7	1.7
29	4	6.0	6.7	10.0	4.0	5	7.5	6.2	9.3	1.8	8	11.9	8.1	12.1	0.2	11	16.4	15.2	22.7	6.3

30	1	1.4	73	10.0	8.6	5	6.8	6.7	9.2	2.4	5	6.8	8.8	12.1	5.3	9	12.3	16.6	22.7	10.4
34	7	13.7	5.1	10.0	-3.7	6	11.8	4.7	9.2	-2.6	2	3.9	6.1	12.0	8.1	6	11.8	11.6	22.7	10.9
38	5	8.8	5.7	10.0	2.2	5	8.8	5.3	9.3	0.5	13	22.8	6.8	12.0	-10.8	13	22.8	13.0	22.8	0.0
41	7	13.7	5.1	10.0	-3.7	6	11.8	4.7	9.2	-2.6	4	7.8	6.1	12.0	4.2	8	15.7	11.6	22.7	7.0
43	7	10.8	6.5	10.0	-0.8	3	4.6	6.0	9.2	4.6	9	13.8	7.8	12.0	-1.8	12	18.5	14.8	22.8	4.3
44	5	6.7	7.5	10.0	3.3	4	5.3	6.9	9.2	3.9	5	6.7	9.0	12.0	5.3	11	14.7	17.1	22.8	8.1
47	4	6.8	5.9	10.0	3.2	3	5.1	5.5	9.3	4.2	5	8.5	7.1	12.0	3.5	14	23.7	13.4	22.7	-1.0
48	22	17.3	12.7	10.0	-7.3	21	16.5	11.7	9.2	-7.3	14	11.0	15.3	12.0	1.0	31	24.4	28.9	22.8	-1.6
49	9	10.2	8.8	10.0	-0.2	9	10.2	8.1	9.2	-1.0	9	10.2	10.6	12.0	1.8	21	23.9	20.0	22.7	-1.2
50	10	13.2	7.6	10.0	-3.2	4	5.3	7.0	9.2	3.9	9	11.8	9.1	12.0	0.2	18	23.7	17.3	22.8	-0.9
51	12	15.2	7.9	10.0	-5.2	9	11.4	7.3	9.2	-2.2	14	17.7	9.5	12.0	-5.7	23	29.1	18.0	22.8	-6.3
55	9	7.0	12.8	10.0	3.0	12	9.4	11.8	9.2	-0.2	7	5.5	15.4	12.0	6.5	27	21.1	29.1	22.8	1.7
57	3	6.0	5.0	10.0	4.0	5	10.0	4.6	9.2	-0.8	6	12.0	6.0	12.0	0.0	12	24.0	11.4	22.8	-1.2
58	7	9.5	7.4	10.0	0.5	4	5.4	6.8	9.2	3.8	6	8.1	8.9	12.0	3.9	19	25.7	16.8	22.7	-3.0
60	8	9.6	8.3	10.0	0.4	5	6.0	7.7	9.3	3.3	9	10.8	10.0	12.0	1.2	22	26.5	18.9	22.8	-3.7
62	8	5.8	13.9	10.0	4.2	8	5.8	12.8	9.2	3.4	14	10.1	16.7	12.0	1.9	23	16.5	31.6	22.7	6.2
66	11	12.5	8.8	10.0	-2.5	14	15.9	8.1	9.2	-6.7	14	15.9	10.6	12.0	-3.9	21	23.9	20.0	22.7	-1.2
68	2	3.8	5.3	10.0	6.2	9	17.0	4.9	9.2	-7.8	13	24.5	6.4	12.1	-12.4	17	32.1	12.1	22.8	-9.3

Table B3

Comparison of Estimated Marginal Outcome Means and Standard Errors by Missing Data Patterns

Grade 2								
Pattern	ORF-F		ORF-W		ORF-S		SAT-10	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
XXXX	38.68	0.72	65.59	0.91	82.67	0.93	66.58	0.44
XXXO	29.35	2.70	51.19	3.41	66.20	3.49	--	--
XXOX	20.00	10.54	34.38	13.32	--	--	55.75	6.46
XXOO	39.83	3.71	68.25	4.69	--	--	--	--
XOXX	35.70	11.68	--	--	74.53	15.07	55.83	7.16
XOXO	16.61	10.21	--	--	44.11	13.18	--	--
XOOO	30.93	2.67	--	--	--	--	--	--
OXXX	--	--	54.53	4.23	68.56	4.32	60.55	2.05
OXXO	--	--	44.62	7.06	61.34	7.21	--	--
OXOX	--	--	23.00	37.68	--	--	52.00	18.28
OXOO	--	--	41.12	6.45	--	--	--	--
OOXX					69.65	7.44	62.72	3.53
OOXO					50.75	6.96	--	--
Grade 3								
	ORF-F		ORF-W		ORF-S		OAKS-Reading	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
XXXX	64.76	0.84	83.01	0.92	100.68	0.91	209.65	0.24
XXXO	54.11	3.14	69.67	3.43	85.85	3.40	--	--
XXOX	45.97	9.59	63.89	10.49	--	--	202.35	2.80
XXOO	51.93	3.80	64.65	4.16	--	--	--	--
XOXX	51.69	11.42	--	--	87.69	12.37	208.44	3.33
XOXO	59.00	15.57	--	--	68.69	16.86	--	--
XOOO	56.49	3.37	--	--	--	--	--	--
XOOX	71.63	16.15	--	--	--	--	208.75	4.71
OXXO	--	--	55.09	7.50	79.00	7.42	--	--
OXOX	--	--	63.75	17.67	--	--	206.25	4.71
OXOO	--	--	79.80	8.17	--	--		
OXXX	--	--	72.81	4.13	89.43	4.09	207.05	1.10
OOXX	--	--	--	--	83.79	7.06	207.02	1.90
OOXO	--	--	--	--	60.52	8.05	--	--

Note. Double dash indicates assessment was not administered. Data was missing by definition.

Table B4

Statistically Significant Comparisons of Missing Error Patterns Using Post-Hoc Bonferroni Corrections

Patterns	OOXO	OOXX	OXOO	OXOX	OXXO	OXXX	XOOO	XOOX	XOXO	XOXX	XXOO	XXOX	XXXO	XXXX
OOXO	-													
OOXX	*DIB-S* 33.93(3)	-												
OXOO			-											
OXOX				-										
OXXO					-									
OXXX	*DIB-S* 29.86(3)					-								
XOOO							-							
XOOX								-						
XOXO		DIBS-S -49.57(3)				DIBS-S -45.51(3)			-					
XOXX										-				
XXOO											-			
XXOX												-		
XXXO	*DIB-F 9.69(3)* *DIB-S* 30.17(3)				*DIB-W* 22.42(3)				*DIB-S 45.81(3)*				-	
XXXX	*DIB-S* 32.99(2) *DIB-S* 41.39(3)		*DIB-W* 25.14(2) *DIB-W* 16.91(3)		*DIB-W* 22.58(2) *DIB-W* 33.79(3) *DIB-S*	*DIB-W* 11.14(3) *DIB-S* 12.47(2) *DIB-S*	*DIB-F* 9.18(2)		*DIB-S* 45.81(2) *DIB-S* 57.04(3)		*DIB-F* 14.52(3) *DIB-W* 18.57(3)		*DIB-F* 8.11(2) *DIB-F* 9.69(3) *DIB-W*	-

24.17(2)	11.53(3)		13.41(2)
	SAT-10 5.61 (2)		*DIB-W* 11.38(3)
DIB-S 26.40(3)			*DIB-S* 16.98(2)
			DIB-S 11.23(3)

Note. *DIB-F* = DIBELS Fall Beginning-of-Year-Score; *DIB-W* = DIBELS Winter Middle-of-Year-Score; *DIB-S* = DIBELS Spring End-of-Year Score.

REFERENCES CITED

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

Algozzine, B., Wang, C., White, R., Cooke, N., Marr, M., Algozzine, K., Duran, G. Z. (2012). Effects of multi-tier academic and behavior instruction on difficult-to-teach students. *Exceptional Children, 79*(1), 45-64.

August, D., & Shanahan, T. (2006). Executive summary. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: A report of the National Literacy Panel on Language-Minority Children and Youth* (pp. 1-9). Retrieved from http://www.cal.org/projects/ archive/natlitpanel.html

Baker, D., Park, Y., & Baker, S. (2012). The reading performance of English learners in grades 1-3: The role of initial status and growth on reading fluency in Spanish and English. *Reading and Writing: An Interdisciplinary Journal, 25*(1), 251-281.

Baker, S. K., Smolkowski, K., Smith, J. M., Fien, H., Kame'enui, E. J. & Beck, C. T. (2011). The impact of Oregon Reading First on student reading outcomes. *The Elementary School Journal, 112*(2), 307-331.

Baker, S. K., Smolkowski, K., Katz, R., Fien, H., Seeley, J. R., Kame'enui, E. J., Beck, C. T. (2008). Reading fluency as a predictor of reading proficiency in low-performing, high-poverty schools. *School Psychology Review, 37*(1), 18-37.

Basaraba, D. L. (2011). *Examining school, student, and measurement effects on first grade students' demonstration of the alphabetic principle* (Doctoral dissertation). Retrieved from Proquest LLC. (UMI No. 3466315)

Berkeley, S., Scruggs, T. E., Mastropieri, M. A. (2010). Reading comprehension instruction for students with learning disabilities, 1995-2006: A meta analysis. *Remedial and Special Education, 31*(6), 423-436.

Berliner, D. (1990). What's all the fuss about instructional time? In M. Ben-Peretz & R. Bromme (Eds.), *The Nature of Time in Schools: Theoretical Concepts, Practitioner Perceptions* (pp. 3-35). New York: Teacher College Press.

Berman, P., & McLaughlin, M. W. (1976). Implementation of educational innovation. *The Educational Forum, 40*(3), 347-370.

Blakely, C. H., Mayer, J. P., Gottschalk, R. G., Schmitt, N., Davidson, W. S., Roitman, D., & Emshoff, J. P. (1987). The fidelity-adaptation debate: Implications for the implementation of public sector social programmes. *American Journal of Community Psychology, 15*(3), 253–268.

Blazer, C. (2007). *Student mobility.* (ERIC Document No. ED541084). Miami, FL: Research Services, Miami-Dade County Public Schools.

Block, C. C., & Duffy, G. G. (2008). Research on teaching comprehension: Where we've been and where we're going. In C. C. Block & S. R. Parris (Eds.), *Comprehension instruction: Research-based best practices* (pp. 19–37). New York: Guilford.

Borman, G. D. (2005). National efforts to bring reform to scale in high-poverty schools: Outcomes and implications. *Review of Research in Education, 29*(1), 1-27.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown (2003). Comprehensive school reform and student achievement: A meta-analysis. *Review of Educational Research, 73*(2), 125-230.

Bradshaw, C. P. Preliminary validation of the implementation phases inventory for assessment fidelity of schoolwide positive behavior supports. *Journal of Positive Behavior Interventions, 11*(3), 145-160.

Buhi, E. R., Goodson, P., & Neilands, T. B. (2008). Out of sight, not out of mind: Strategies for handling missing data. *American Journal of Health Behavior, 32*(1), 83-92.

Brandon, P. R., Lawton, B. E., & Harrison, G. M. (2014). Issues of rigor and feasibility when observing the quality of program implementation: A case study. *Evaluation and Program Planning, 44,* 75-80.

Bryk, A. S., & Raudenbush, S. W. (1992). Hierarchical linear models: Applications and data analysis methods. Newbury Park: Sage.

Carlisle, J. F., & Berebitsky, D. (2011). Literacy coaching as a component of professional development. *Reading and Writing, 24*, 773-800.

Carlisle, J. F., Cortina, K. S., & Zeng, J. (2010). Reading achievement of Reading First schools in Michigan. *Journal of Literacy Research, 42*(1), 49-70.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*(8), 723-733.

Castro, F., Barrera Jr., M., & Steiker, L. (2010). Issues and challenges in the design of culturally adapted evidence-based interventions. *Annual Review of Clinical Psychology, 6*, 213-239.

Century, J. Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation, 31*(2), 199-218.

Chard, D., Ketterlin-Geller, L., Baker, S., Doabler, C., & Apichatabutra, C. (2009). Repeated reading interventions for students with learning disabilities: Status of the evidence. *Exceptional Children, 75*(3), 263-281.

Cheatham, J. P., & Allor, J. H. (2012). The influence of decodability in early reading text on reading achievement: A review of the evidence. *Reading and Writing, 25*(9), 2223-2246.

Christ, T. J., Zopluoglu, C., Long, J. D., & Monaghen, B. D. (2012). Curriculum-based measurement of oral reading: Quality of progress monitoring outcomes. *Exceptional Children, 78*(3), 356-373.

Coburn, C. E., & Woulfin, S. L. (2012). Reading coaches and the relationship between policy and practice. *Reading Research Quarterly, 47*(1), 5-30.

Colorado Department of Education RTI/PBIS Unit. (2011*). RtI implementation rubrics guidebook*. Denver, CO: Colorado Department of Education.

Connor, C. M., Jakobsons, L. J., Crowe, E. C., & Meadows, J. G. (2009). Instruction, student engagement, and reading skill growth in Reading First classrooms. *The Elementary School Journal, 109*(3), 221-250.

Cook, B. G., & Odom, S. L. (2013). Evidence-based practice and implementation in special education. *Exceptional Children, 79*(2), 135-144.

Cook, R. C., Mayer, G. R., Wright, D. B., Kraemer, B., Wallace, M. D., Dart, E., . . . Restori, A. (2012). Exploring the link among behavior intervention plans, treatment integrity and student outcomes under natural educational conditions. *The Journal of Special Education, 46*(1), 3-16.

Correnti, R., & Rowan, B. (2007). Opening up the black box: Literacy instruction in schools participating in three comprehensive school reform programs. *American Educational Research Journal, 44*(2), 298-338.

Cotton, K. (2001). Using school time productively. In J. S. Chick (Ed.), *Traditional school*. Kansas City, MO: School Improvement Research Series Snapshot #22.

Crawford, L., Carpenter II, D. M., Wilson, M. R., Schmeister, M., & McDonald, M. (2012). Testing the relation between fidelity of implementation and student outcomes in math. *Assessment for Effective Intervention, 37*(4), 224-235.

Creighton, T. B. (2001). Data analysis and the principalship. *Principal Leadership, 1*(9), 52-57.

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology*, *33*, 934–945.

Dadnow, A., Park, V., & Wohlstetter (2007). *Achievement with Data: How high-performing school systems use data to improve instruction for elementary students.* Center on Educational Governance, University of Southern California: New Schools Venture Fund.

Darling-Hammond, L., & Richardson, N. (2009). Teacher learning: What matters? *Educational Leadership, 66*(5), 46-53.

De Fazio, C. M., Fain, A. C., & Duchaine, E. L. (2011). Using treatment integrity in the classroom to bring research and practice together. *Beyond Behavior, 20*, 45-49.

Denton, C. A. (2012). Response to intervention for reading difficulties in the primary grades: Some answers and lingering questions. *Journal of Learning Disabilities, 45*(3), 232-243.

Denton, C. A., Vaughn, S., Tolar, T. D., Fletcher, J. M., Barth, A. E. & Francis, D. J. (2013). Effects of tier 3 intervention for students with persistent reading difficulties and characteristics of inadequate responders. *Journal of Educational Psychology, 105*(3), 633-648.

Dewitz, P., Jones, J., & Leahy, S. (2009). Comprehension strategy instruction in core reading programs. *Reading Research Quarterly, 44*(2), 102–126.

Dexter, D.D., & Hughes. (2009). Progress Monitoring and identifying non-responders in Tier 2. RTI Action Network Web, National Center for Learning Disabilities. Retrieved from http://www.rtinetwork.org/Learn/Research/ar/ResearchReview

Dobbie, W., & Fryer, R, Jr. (2011). *Getting beneath the veil of effective schools: Evidence from New York City.* (EBER Working Paper, No. 17632). Cambridge, MA: National Bureau of Economic Research.

Domenech Rodríguez, M., Baumann, A., & Schwartz, A. (2011). Cultural adaptation of an evidence based intervention: From theory to practice in a latino/a community context. *American Journal of Community Psychology*, *47*(1-2), 170.

Drake, R., Goldman, H., Leff, H., Lehman, A., Dixon, L., Mueser, K., & Torrey, W. (2001). Implementing evidence-based practices in routine mental health service settings. *Psychiatric Services, 52*, 179-182.

Duffy, G. G., Roehler, L. R., Meloth, M. S., Vavrus, L. G., Book, C., Putnam, J., & Wesselman, R. (1986). The relationship between explicit verbal explanations during reading skill instruction and student awareness and achievement: A study of reading teacher effects. *Reading Research Quarterly, 21*(3), 237–252.

DuFour, R. (2007). Professional learning communities: A bandwagon, an idea worth considering, or our best hope for high levels of learning? *Middle School Journal*, 39, 4–8.

DuFour, R., & Marzano, R. (2009). High leverage strategies for principal leadership. *Educational Leadership, 66*(5), 62-68.

Duke, N. K., & Pearson, P. D. (2002). Effective practices for developing reading comprehension. In A. E. Farstrup & S. J. Samuels (Eds.), *What research has to say about reading instruction* (pp. 205–242). Newark, DE: International Reading Association.

Dumas, J., Lynch, A., Laughlin, J., Smith, E., & Prinz, R. (2001). Promoting intervention fidelity: Conceptual issues, methods and preliminary results from the EARLY ALLIANCE prevention trial. *American Journal of Preventive Medicine, 20*(1 Suppl.), 38–47.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*(3-4), 327-350.

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research Theory and Practice, 18*, 237-256.

Dynamic Measurement Group (2008). DIBELS 6th Edition Technical Adequacy Information (Tech. Rep. No. 6). Eugene, OR: Author.

Eaker, R., & Keating, J. (2008). A shift in school culture: Collective commitments focus on change that benefits student learning. *Journal of Staff Development*, 29, 14–17.

Edmonds, M., Vaughn, S., Hjelm, J., Reutebuch, C., Cable, A., & Tackett, K., & Schnakenberg, J. W. (2009). A synthesis of reading interventions and effects on reading outcomes for struggling readers. *Review of Educational Research, 79*, 266–300.

Ehri, L. C. (2003, March*). Systematic phonics instruction: Findings of the National Reading Panel.* Paper presented at the Invitational Seminar of the Standards and Effectiveness Unit, Department for Education and Skills, British Government. London, England.

Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading, 9*(2), 167-188.

Elleman, A. M., Lindo, E. J., Morphy, P., & Compton, D. L. (2009). The impact of vocabulary instruction on passage level comprehension of school age children: A meta-analysis. *Journal of Research on Educational Effectiveness, 2,* 1-44.

Elmore, R. F. (2002). *Bridging the Gap Between Standards and Achievement: Report on the Imperative for Professional Development in Education.* Washington, DC: Albert Shanker Institute.

Elmore, R. F. (1998, November). *The National Education Goals Panel: Purposes, Progress, and Prospects*. Paper prepared for the National Education Goals Panel. Boston, MA: Harvard University.

Elmore, R. F. (2004). *School reform from the inside out: Policy, practice, and performance.* Cambridge, MA: Harvard Education Press.

Ennett, S. T., Haws, S., Ringwalt, C. L., Vincus, A. A., Hanley, S., Bowling, J. M., & Rohrbach, L. A. (2011). Evidence-based practice in school substance use prevention: Fidelity of implementation under real-world conditions. *Health Education Research, 26*(2), 361-371.

Erickson, A. G., Noonan, P. M., & Jenson, R. (2012). The school implementation scale: Measuring implementation in response to intervention models. *Learning Disabilities: A Contemporary Journal, 10*(2), 33-52.

Fagan, A. A., Hanson, K., Hawkins, J. D., & Arthur, N. W. (2008). Implementing effective community-based prevention programs in the community youth development study. *Youth Violence and Juvenile Justice, 6,* 256-278.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M. & Wallace, F. (2005). *Implementation Research: A Synthesis of the Literature.* (FMHI Publication #231). Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network.

Foorman, B., & Torgesen, J. (2001). Critical elements of classroom and small-group instruction promote reading success in all children. *Learning Disabilities Research and Practice, 16*(4), 203-212.

Fox, D. (2013). The principal's mind-set for data. *Leadership, 42*(3), 12-36.

Freebody, P., & Anderson, R. C. (1983). Effects of text comprehension of differing proportions and locations of difficult vocabulary. *Journal of Reading Behavior*, 15, 19–39.

Fuchs, L. S., & Fuchs, D. (2003). *What is scientifically-based research on progress monitoring?* Retrieved from http://www.studentprogress.org/library/articles.asp#whatisresearch

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical and historical analysis. *Scientific Studies of Reading, 5*(3), 239-256.

Fuchs, L. S., & Vaughn, S. (2012). Responsiveness-to-intervention: A decade later. *Journal of Learning Disabilities, 45*(3), 195-203.

Fullan, M., Cuttress, C., Kilcher, A. (2005). 8 forces for leaders of change. *Journal of Staff Development, 26*, 54-64.

Fullan, M., & Pomfret, A. (1977). Research on curriculum and instruction implementation *Review of Educational Research, 47*(2), 335-397.

Gamse, B. C., Jacob, R. T., Horst, M., Boulay, B., and Unlu, F. (2008). *Reading First Impact Study Final Report* (NCEE 2009-4038). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gansle, K. A., & Noell, G. H. (2007). The fundamental role of intervention implementation in assessing response to intervention. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden. (Eds.), *Handbook of response to intervention: The science and practice of assessment and intervention* (pp. 244 – 251). New York: Springer.

Garet, S., Porter, A., Desimone, L., Birman, B., & Yoon, K. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*, 915-945.

Gersten, R., Compton, D., Connor, C. M., Dimino, J., Santoro, L., Linan-Thompson, S., & Tilly, W. D. (2008). *Assisting students struggling with reading: Response to intervention and multitier intervention for reading in the primary grades. A practice guide* (NCEE 2009–4045). Washington, DC: U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences. Retrieved from http://ies.ed.gov/ncee/wwc/publications/practiceguides/

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

Gilbert, J. K., Compton, D. L., Fuchs, D., & Fuchs, L. S. (2012). Early screening for risk of reading disabilities: Recommendations for a four-step screening system. *Assessment for Effective Intervention, 38*(1), 6-14.

Gilbert, J. K., Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Barquero, L. A., Cho, E. (2013). Responsiveness-to-intervention prevention model for struggling readers. *Reading Research Quarterly, 48*(2), 135-154.

Good, R. H. & Kaminski, R. A. (2002). *DIBELS Oral Reading Fluency Passages for First through Third Grades* (Technical Report No. 10). Eugene, OR: University of Oregon.

Good III, R. H., Simmons, D. C., & Smith, S. B. (1998). Effective academic interventions in the United States: Evaluating and enhancing the acquisition of early reading skills. *School Psychology Review, 27*(1), 45-57.

Glover, T. A., & DiPerna, J. C. (2007). Service delivery for response to intervention: Core components and directions for future research. *School Psychology Review, 36*(4), 632-637.

Gottfried, M. (2009). Excused versus unexcused: How student absences in elementary school affect academic achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 392-415.

Greenwood, C. (1991). Longitudinal analysis of time, engagement and achievement in at-risk versus non-risk students. *Exceptional Children, 57,* 521-535.

Grigg, J. (2012). School enrollment changes and student achievement growth: A case study in educational disruption and continuity. *Sociology of Education, 85*(4), 388-404.

Griffiths, A., Parson, L. B., Burns, M. K., VanDerHeyden, A., Tilly, W. D. (2007). *Response to intervention: Research for practice*. National Association of State Directors of Special Education, Inc. Alexandria, VA: Author.

Griffiths, A., VanDerHeyden, A. M., Parson, L. B., & Burns, M. K. (2008). Practical applications of response-to-intervention research. *Assessment for Effective Intervention, 32*(1), 50-57.

Griner, D., & Smith T.B. (2006). Culturally adapted mental health interventions: A meta-analytic review. *Psychotherapy: Theory, Research, Practice, Training, 43*(4), 531–48.

Grossen, B. (1997). *Thirty years of research: What we know about how children learn to read: A synthesis of research on reading from the National Institute of Child Health and Human Development*. Santa Cruz, CA: The Center for the Future of Teaching and Learning.

Harcourt Educational Measurement. (2002). *Stanford Achievement Test [SAT-10]*. San Antonio,
TX: Author.

Harms, A. (2010). A three-tier model of integrated behavior and learning supports: Linking system-wide implementation to student outcomes (Doctoral dissertation). Retrieved from ProQuest LLC. (UMI No. 3433110).

Harn, B., Parisi, D., & Stoolmiller, M. (2013). Flexibility and fit: What do we really know about fidelity of implementation in schools? *Exceptional Children, 79*(2), 181-193.

Hart, B., & Risley, T. R. (1995). *Meaningful Differences in the Everyday Experiences of Young American Children*. Baltimore: Paul H. Brookes.

Hilton, A. (2007). Response to Intervention: Changing how we do business. *Leadership, 3,* 16-19.

Hudson, R., Torgesen, J., Lane, H., & Turner, S. (2012). Relations among reading skills and sub-skills and text-level reading proficiency in developing readers. *Reading and Writing, 25*(2), 483-507.

Hughes, C. A., & Dexter, D. D. (2011). Response to Intervention: A research-based summary. *Theory Into Practice,* 50(1), 4-11.

Hunley, S., Davies, S., & Miller, C. (2013). The relationship between curriculum-based measures in oral reading fluency and high-stakes tests for seventh grade students. *RMLE Online, 36*(5), 1-8.

Jenkins, J. R., Hudson, R. H., & Johnson, E.S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36*(4), 582-600.

Jeynes, W. H. (2008). A meta-analysis of the relationship between phonics instruction and minority elementary school student academic achievement. *Education and Urban Society, 40*(2), 151-166.

Kame'enui, E. J., & Simmons, D. C. (2003). *Planning and Evaluation Tool for Effective School-wide Reading Programs- Revised (PET-R).* Eugene, OR: Institute for the Development of Educational Achievement.

Kame'enui, E. J., Simmons, D. C., & Coyne, M. D. (2000). Schools as host environments: Toward a schoolwide reading improvement model. *Annals of Dyslexia, 50,* 33-51.

Kamil, M., & Hiebert, E. (2005). Teaching and learning vocabulary: Perspectives and persistent issues. In E. H. Hiebert and M. L. Kamil (Eds.*), Teaching and learning vocabulary: Bringing research to practice* (pp. 1-23). Mahwah, NJ: Lawrence Erlbaum.

Kamps, D., Abbott, M., Greenwood, C., Wills, H., Veerkamp, M., et al. (2008). Effects of small-group reading instruction and curriculum differences for students most at risk in kindergarten: Two-year results for secondary- and tertiary-level interventions. *Journal of Learning Disabilities, 41*(2), 101-114.

Kansas Multi-Tier System of Supports (2010). *The integration of MTSS and RtI.* Author. Retrieved from http://www.kansasmtss.org/briefs/The_Integration_of_MTSS_and_RtI.pdf

Kelleher, J. (2003). Professional development that works: A model for assessment-driven professional development. *Phi Delta Kappan, 84,* 751-756.

Keller-Margulis, M. A. (2012). Fidelity of implementation framework: A critical need for response to intervention models. *Psychology in the Schools, 49*(4), 342-352.

Kendall, P. C., Gosch, E., Furr, J. M., & Sood, E. (2008). Flexibility within fidelity. *Journal of the American Academy of Child and Adolescent Psychiatry, 47,* 987–993.

Kim, W., Linan-Thompson S., & Misquitta, R. (2012). Critical factors in reading comprehension for students with learning disabilities: A research synthesis. *Learning Disabilities Research & Practice, 27*(2), 66-78.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of discourse comprehension and production. *Psychological Review, 83,* 363-394.

Kitchens, J. (2005). *Real-time access to student data leads to real school reform*. eSchool News Online: 3. Retrieved from http://www.eschoolnews.com/2005/10/01/real-time-access-to-student-data-leads-to-real-school-reform.

Kovaleski, J. F. (2007). Response to intervention: Considerations for research and systems change. *School Psychology Review, 36,* 638-646.

Kovaleski, J. F., Marco-Fies, C. M., & Boneshefski, M. J. (2013). Treatment Integrity: Ensuring the "I" in RtI. *RTI Action Network.* Retrieved from http://rtinetwork.org/getstarted/evaluate/treatment-integrity-ensuring-the-i-in-rti

Kratochwill, T. R., Volpiansky, P., Clements, M., & Ball, C. (2007). Professional development in implementing and sustaining multitier prevention models: Implications for response to intervention. *School Psychology Review, 36,* 618-631.

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

Lachat, M. A., Williams, M., & Smith, S. C. (2006). Making sense of all your data. *Principal Leadership,7*(2), 16-21.

LeMahieu, P. (2011) What we need in education is more integrity (and less fidelity) of implementation. *R & D Ruminations.* Carnegie Foundation.

Lendrum, A., & Humphrey, N. (2012). The importance of studying the implementation of interventions in school settings. *Oxford Review of Education, 38*(5), 635-652.

Lesaux, N., & Kieffer, M. (2010). Exploring sources of reading comprehension difficulties among language minority learners and their classmates in early adolescence. *American Educational Research Journal, 47*(3), 596-632.

Levin, J., & Datnow, A. (2012). The principal role in data-driven decision making: Using case-study data to develop multi-mediator models of educational reform. *School Effectiveness and School Improvement, 23*(2), 179-201.

Little, R. J. A. (1998) A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*, 1198-1202.

Logan, J. A. R., & Petscher, Y. (2010). School profiles of at-risk student concentration: Differential growth in oral reading fluency. *Journal of School Psychology, 48*(2), 163-186.

Lomos, C., Hofman, R. H., & Bosker, R. J. (2011). Professional communities and student achievement – a meta-analysis. *School Effectiveness and School Improvement, 22*(2), 121-148.

Lynch, S., & O'Donnell, C. (2005, April). The evolving definition, measurement, and conceptualization of fidelity of implementation in scale-up of highly rated science curriculum units in diverse middle schools. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Mandinach, E. B., Honey, M., & Light, D. (2006, April). A theoretical framework for data-driven decision making. Paper presented at the Annual Meeting of American Educational Research Association, San Francisco, California.

Margolis, H. (2012). Response to intervention: RTI's linchpins. *Reading Psychology, 33*(1-2), 8-10.

Marston, D., Mirkin, P., & Deno, S. (1984). Curriculum-based measurement: An alternative to traditional screening, referral, and identification. *Journal of Special Education, 18*(2), 109-117.

Maynard, B. R., Peters, K. E., Vaughn, M. G., & Sarteschi, C. M. (2013). Fidelity in after-school program intervention research: A systematic review. *Research on Social Work Practice, 23*(6), 613-623.

McArthur, G., Eve, P. M., Jones, K., Banales, E., Kohnen, S., Anandakumar, T., . . . Castles, A. (2012). Phonics training for English-speaking poor readers. *Cochrane Database of Systematic Reviews,* Issue 12.

McGlinchey, M. T., & Hixson, M. D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review, 33*(2), 193-203.

McKenna, M. K., & Good, R. H., III. (2003). *Assessing reading comprehension: The relation between DIBELS Oral Reading Fluency, DIBELS Retell Fluency, and Oregon State Assessment scores.* Eugene: University of Oregon.

McKenna, M. C., & Walpole, S. (2010). Planning and evaluating change at scale: Lessons from Reading First. *Educational Researcher, 39*(6), 478-483.

McLeod, B. D., Southam-Gerow, M. A., & Weisz, J. R. (2009). Conceptual and methodological issues in treatment integrity measurement. *School Psychology Review, 38*(4), 541-546.

McMaster, K. L., Espin, C. A., & van den Broek (2014). Making connections: Linking cognitive psychology and intervention research to improve comprehension of struggling readers. *Learning Disabilities Research & Practice, 29*(1), 17-24.

McMaster, K. L., van den Broek, P., Espin, C. A., White, M. J., Rapp, D. N., Kendeou, P., Bohn-Gettler, C. M., & Carlson, S. (2012). Making the right connections: Differential effects of reading intervention for subgroups of comprehenders. *Learning and Individual Differences, 22*(1), 100-111.

Mehana, M, & Reynolds, A. (2004). School mobility and achievement: A meta-analysis. *Children and Youth Services Review, 26*(1), 93-119.

Melby-Lervag, M., & Lervag, A. (2014). Effects of educational interventions targeting reading comprehension and underlying components. *Child Development Perspectives, 8*(2), 96-100.

Melby-Lervag, M., Lyster, S. H., & Hulme, C. (2012). Phonological skills and their role in learning to read: A meta-analytic review. *Psychological Bulletin, 138*(2), 322-352.

Mellard, D. F., Frey, B. B., & Woods, K. L. (2012). School-wide student outcomes of response to intervention frameworks. *Learning Disabilities: A Contemporary Journal, 10*(2), 17-32.

Mellard, D. F., Prewett, S., & Deshler, D. (2012). Strong leadership for RTI success. *Principal Leadership, 12*(8), 28-32.

Mihalic, S. F., & Irwin, K. (2003). Blueprints for violence prevention: From research to real world settings. Factors influencing the successful replication of model programs. *Youth Violence and Juvenile Justice, 1*, 1–23.

Moore, J. E., Bumbarger, B. K., Cooper, B. R. (2013). Examining adaptations of evidence-based programs in natural contexts. *Journal of Primary Prevention, 34*(3), 147-161.

Morrison, D. (2005) Principal leadership competencies in a successful school reform effort. (Doctoral Dissertation). Retrieved from ProQuest LLC. (UMI No. 3222538).

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24*(3), 315-340.

Nagy. W. E. (2007). Metalinguistic awareness and the vocabulary comprehension connection. In R. Wagner, A. Muse, & K. Tannenbaum (Eds.). *Vocabulary acquisition: Implications for reading comprehension* (pp. 53-77). New York: Guilford Press.

National Assessment Governing Board. (2006). *Reading framework for the 2007 National Assessment of Educational Progress.* Washington, DC: U. S. Department of Education.

National Center for Learning Disabilities. (2012). *Issue Brief: Multi-tier system of supports aka Response to Intervention (RTI).* Washington, DC: Author.

National Center on Response to Intervention. (2011). RTI Essential Components Integrity Rubric. Washington, DC: U.S. Department of Education, Office of Special Education. Programs, National Center on Response to Intervention.

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

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *The Journal of Behavioral Health Services & Research, 39*(4), 374-396.

NCLD Public Policy Team. (2013). *Multi-tier System of Support aka Response to Intervention (RtI). Issue Brief: Multi-Tier System of Supports (MTSS) Recommendation for the Elementary and Secondary Education Act (ESEA).*

National Center for Learning Disabilities. Retrieved from http://www.ncld.org/disability-advocacy/where-we-stand-policies/multi-tier-system-supports-response-intervention.

NICHD Early Child Care Research Network. (2005). Pathways to reading: The role of oral language in the transition to reading. *Developmental Psychology, 41*(2), 428–442.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002). U. S. Department of Education (2002). Guidance to the Reading First Program.

Noell, G. H., Witt, J. C., Slider, N. J., Connell, J. E., Gatti, S. L., Williams, K. L., Koenig, J. L., & Resetar, J. L. (2005). Treatment implementation following behavioral consultation in schools: A comparison of three follow-up strategies. *School Psychology Review, 34*(1), 87-106.

Noltemeyer, A. L., Boone, W. J., & Sansosti, F. J. (2014). Assessing school-level RTI implementation for reading: Development and piloting the RTIS-R. *Assessment for Effective Intervention.*

O'Connor, R., Harty, K., & Fulmer, D. (2005). Tiers of intervention in kindergarten through third grade. *Journal of Learning Disabilities*, *38*(6), 532-538.

Odom, S. L., Fleming, K., Diamond, K.., Lieber, J., Hanson, M., Butera, G., Horn, E., Palmer, S., & Marquis, J. (2010). Examining different forms of implementation and in early childhood research. *Early Childhood Research Quarterly, 25* (3), 324-328.

Odom, S. L. (2009). The ties that bind: Evidence based practice, implementation science, and outcomes for children. *Topics in Early Childhood Special Education, 29*, 53–61.

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research, 78*(1), 33-84.

Ogden, T., & Fixsen, D. L. (2014). Implementation science: A brief overview and a look ahead. *Zeitschrift fur Psychologie, 222*(1), 4-11.

Olson, L. (2003). NAEP worries states excluding too many from tests. *Education Week, 22*(27), 7.

Oregon Department of Education Office of Assessment and Information Services (2004). Knowledge and Skills Assessment Manual 2004-05 School Year. Salem, OR: Author.

Oregon Department of Education (2007). 2006-2007 Technical Report – Reliability and Validity, Salem, OR: Author.

Parisi, D. (2009). Examining multiple dimensions of fidelity and their relation to student reading outcomes: A retrospective analysis of kindergarten interventions (Doctoral Dissertation). Retrieved from ProQuest LLC. (UMI No. 3377388).

Panettieri, J. S. (2006). data DRIVEN. *T.H.E. Journal, 33*(7), 24-29.

Parker, L. (2005). The elementary and secondary education act at 40: Reviews of research, policy Implementation, critical perspectives, and reflections. *Review of Research in Education, 29*.

Pas, E. T., & Bradshaw, C. P. (2012). Examining the association between implementation and outcomes. *Journal of Behavioral Health Services & Research, 39*(4), 417-433.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions. *Review of Educational Research, 74*(4), 525-556.

Pinnell, G. S., Pikulski, J. J., Wixson, K. K., Campbell, J. R., Gough, P. B., & Beatty, A. S. (1995). *Listening to children read aloud.* Washington, DC: Office of Educational Research and Improvement, U. S. Department of Education.

Porche, M. V., Pallante, D. H., & Snow, C. E. (2012). Professional development for teaching reading achievement: Results from the Collaborative Language and Literacy Instruction Project (CLLIP). *Elementary School Journal, 112*(4), 649-671.

Quirk, M., & Beem, S. (2012). Examining the relations between reading fluency and reading comprehension for English language learners. *Psychology in the Schools, 49*(6), 539-553.

Raudenbush, S.W., Bryk, A.S, & Congdon, R. (2013). HLM 7 for Windows [Computer software]. Skokie, IL: Scientific Software International, Inc.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage.

Reschly, A., Busch, T., Betts, J., Deno, S., & Long, J. (2009). Curriculum-based measurement of oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47*(6), 427-469.

Roberts, J. K. (2007, April). *Group dependency in the presence of small intraclass correlation coefficients: An argument in favor of not interpreting the ICC*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Illinois.

Rowan, B., Camburn, E., & Barnes, C (2004). Benefiting from comprehensive school reform: A review of research on CSR implementation. In C. Cross (Ed*.), Putting the pieces together: Lessons from comprehensive school reform research* (pp. 1-52). Washington, DC: National Clearinghouse for Comprehensive School Reform.

Sanetti, L. M. H., Gritter, K. L., & Dobey, L. M. (2011). Treatment integrity of interventions with children in the school psychology literature from 1995 to 2008. *School Psychology Review, 40*(1), 72-84.

Sanettei, L. M. H., & Kratochwill, T. R. (2011). An evaluation of the treatment integrity planning protocol and two schedules of treatment integrity self-report: Impact on implementation and report accuracy. *Journal of Education and Psychological Consultation, 21,* 284-308.

Sanettei, L. M. H., & Kratochwill, T. R. (2009). Toward developing a science of treatment integrity: Introduction to the special series. *School Psychology Review, 39*(4), 445-459.

Sanford, A., Park, Yonghan, Baker, S. K. (2013). Reading growth of students with disabilities in the context of a large-scale statewide reading reform effort. *Journal of Special Education, 47*(2), 83-95.

Scarborough, H. S. (2001). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In S. Neuman & D. Dickinson (Eds.), *Handbook for research in early literacy* (pp. 97-110). New York: Guilford Press.

Schmitt, N., Jiang, X., & Trabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 9*(11), 26-43.

Shadish, W. R. (2011). Randomized controlled studies and alternative designs in outcome studies: Challenges and opportunities. *Research on Social Work Practice, 21*(6), 636-643.

Shadish, W. R., & Ragsdale, K. (1996). Random versus nonrandom assignment in controlled experiments: Do you get the same answer*? Journal of Consulting and Clinical Psychology, 64*(6), 1290-1305.

Shanahan, T. (2005). *The National Reading Panel Report: Practical Advice for Teachers.* Learning Points Associates. Naperville, IL.

Shapiro, E., & Clemens, N. (2009). A conceptual model for evaluating system effects of response to intervention. *Assessment for Effective Intervention, 35*(1), 3-16.

Sheppard, B. (2008). The effect of noninstructional workload tasks upon instructional time for classroom teachers in public schools in an urban school district (Doctoral dissertation). Retrieved from ProQuest LLC. (UMI No. 3319380).

Sheridan, S. M., Swanger-Gagne, M., Welch, G. W., Dwon, K., & Garbacz, S. A. (2009). Fidelity measurement in consultation: Psychometric issues and preliminary examination. *School Psychology Review, 38,* 476-495.

Shin, M. R. (2007). Identifying students at risk, monitoring performance, and determining eligibility within response to intervention: Research on educational need and benefit from academic intervention. *School Psychology Review, 36*(4), 601-617.

Shinn, M. R. (2008). Best practices in using curriculum-based measurement in a problem-solving model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp.243-262). Bethesda, MD: National Association of School Psychologists.

Siegel, L. (1993). Phonological processing deficits as the basis of a reading disability. *Developmental Review, 13*(2), 246-257.

Simmons, D., Kame'enui, E., Harn, B., Coyne, M., Stoolmiller, M., Santoro, L. E., . . . Kaufman, N. K. (2007). Attributes of effective and efficient kindergarten reading intervention: An examination of instructional time and design specificity. *Journal of Learning Disabilities, 40*(4), 331-347.

Simmons, D. C., Kuykendall, K., King, K., Cornachione, C., & Kame'enui, E. J. (2000). Implementation of a schoolwide reading improvement model: "No one ever told us it would be this hard!" *Learning Disabilities Research & Practice, 15*(2), 92-100.

Snow, C. E. (2002). *Reading for understanding: Toward an R & D program in reading comprehension.* Santa Monica, CA: RAND.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children.* Committee on the Prevention of Reading Difficulties in Young Children. Washington, DC: National Research Council.

Sparks, R. L., Patton, J., Murdoch, A. (2014). Early reading success and its relationship to reading achievement and reading volume: Replication of '10 years later.' *Reading and Writing, 27*(1), 189-211.

Spear-Swerling, L., & Cheesman, E. (2012). Teachers' knowledge base for implementing response-to-intervention models in reading. *Reading and Writing: An Interdisciplinary Journal, 25*(7), 1691-1723.

Stahl, K., & McKenna, M. (2006*). Reading research at work: Foundations of effective practice*. New York: Guilford Press.

Stahl, S. A., & Murray, B. A. (1994). Defining phonological awareness and its relationship to early reading. *Journal of Educational Psychology, 86*(2), 221-234.

Stahl, S. A., & Fairbanks, M. M. (2006). The effects of vocabulary instruction – A model-based meta analysis. *Review of Educational Research, 56*(1), 72-110.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2007). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools, 42*(8), 795-819.

Stecker, P. M, Fuchs, D., Fuchs, L. S. (2008). Progress monitoring as essential practice within response to intervention. *Rural Special Education Quarterly, 27*(4), 10-17.

Stein, M. L., Berends, M., Fuchs, D., McMaster, K., Saenz, L., Yen, L., Fuchs, L. S., & Compton, D. L. (2008). Scaling up an early reading program: Relationships among teacher support, fidelity of implementation, and student performance across different sites and years. *Educational Evaluation and Policy Analysis, 30*(4), 368-388.

Steiner, P. M., Wroblewski, A., & Cook, T. D. (2009). Chapter 5. Randomized experiments and quasi-experimental designs in education research. In K. E. Ryan & J. B. Cousins (Eds.), *The SAGE international handbook of education evaluation*, pp. 75-96. Thousand Oaks, CA: Sage Publications, Inc.

Stockslager, K., Castillo, J., Brundage, A., Childs, K., & Romer, N. (2014, February). *Developing a school-level tool to monitor implementation of MTSS*. Presentation at the 2014 National Association of School Psychologists Annual Convention. Florida Multi-Tiered System of Supports.

Sullivan, G., Blevins, D., & Kauth, M. R. (2008). Translating clinical training into practice in complex mental health systems: Toward opening the ''Black Box'' of implementation. *Implementation Science, 3*, 33.

Supovitz, J. A., & Klein, V. (2003). *Mapping a course for improved student learning: How innovative schools systematically use student performance data to guide improvement.* Consortium for Policy Research in Education: University of Pennsylvania.

Swanson, E. A. (2008). Observing reading instruction for students with learning disabilities: A synthesis. *Learning Disability Quarterly, 31*(3), 115-133.

Swanson, E., Wanzek, J., Haring, C., Ciullo, S., & McCulley (2013). Intervention fidelity in special and general education. *The Journal of Special Education, 47*(1), 3-13.

Tannenbaum, C. (2009). *The empirical nature and statistical treatment of missing data* (Doctoral dissertation). Retrieved from ProQuest LLC. (UMI No. 13381876).

Therrien, W. (2004). Fluency and comprehension gains as a result of repeated reading: A meta-analysis. *Remedial and Special Education, 25*(4), 252-261.

Thompson, S. (2011). *School size, school poverty and school-level mobility: Interactive threats to school outcomes*. ProQuest. (UMI No. 3447911)

Tilly, W. D., Harken, S., Robinson, W., & Kurns, S. (2008). 3 tiers of intervention. *School Administrator, 65*(8), 20-23.

Torgesen, J. K., & Mathes, P. G. (2000). *A basic guide to understanding, assessing, and teaching phonological awareness.* Austin, TX: Pro-Ed.

U.S. Department of Education, Office of Planning, Evaluation, and Policy Development (2010). *Use of Education Data at the Local Level From Accountability to Instructional Improvement*, Washington, D.C.: Author.

U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service (2011). *Reading First Implementation Study 2008–09 Final Report*, Washington, D.C.: Author.

VanDerHeyden, A., McLaughlin, T., Albina, J., & Snyder, P. (2012). Randomized evaluation of a supplemental grade-wide mathematics intervention. *American Educational Research Journal, 49*(6), 1251-1284.

VanDerHeyden, A. M., Witt, J. C., & Gilbertson, D. (2007). A multiyear evaluation of the effects of a response to intervention (RTI) model on identification of children for special education. *Journal of School Psychology*, *45*, 225–256.

VanDerHeyden, A. M., Witt, J. C., Naquin, G., & Noell, G. (2001). The reliability and validity of curriculum-based readiness probes for kindergarten students. *School Psychology Review, 30*(3), 363-382.

Vaughn, S. (2014). Intensive interventions in reading for students with reading disabilities: Meaningful impacts. *Learning Disabilities Research & Practice, 29*(2), 46-53.

Verhoeven, L., & Van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology, 22*(3), 407-423.

Vogel, L. R. (2010). *Leading Standards-Based Education Reform: Improving Implementation of Standards to Increase Student Achievement.* Rowman & Littlefield Education. Eric Document ED533222.

Wagner, R. K., & Meros, D. (2010). Vocabulary and reading comprehension: Direct, indirect, and reciprocal influences. *Focus on Exceptional Children, 43*(1), 1-10.

Wanzek, J., Al Otaiba, S., & Petscher, Y. (2014). Oral reading fluency development for children with emotional disturbance or learning disabilities. *Exceptional Children, 80*(2), 187-204.

Wanzek, J., Roberts, G., Linan-Thompson, S., Vaughn, S., Woodruff, A, & Murray, C. S. (2010). Differences in the relationship of oral reading fluency and high-stakes measures of reading comprehension. *Assessment for Effective Intervention, 35*(2), 67-77.

Wanzek, J., & Vaughn, S. (2007). Research-based implications from extensive early reading interventions. *School Psychology Review, 36*(4), 541-561.

Wanzek, J., & Vaughn, S. (2010). Tier 3 interventions for students with significant reading problems. *Theory Into Practice, 49,* 305-314.

Wayman, M., Wallace, T., Wiley, H. L., Ticha, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *Journal of Special Education, 41,* 85-120.

What Works Clearinghouse. (2012*). Phonological awareness training.* What Works Clearinghouse Intervention Report. *What Works Clearinghouse.*

Wilson, A. M., & Lesaux, N. (2001). Persistence of phonological processing deficits in college students with dyslexia who have age-appropriate reading skills. *Journal of Learning Disabilities, 34*(5), 394-400.

Wise, J., Sevcik, R., Morris, R., Lovett, M., Wolf, N., Kuhn, M., Beisinger, B., & Schwanenflugel, P. (2010). The relationship between different measures of oral reading fluency and reading comprehension second-grade students who evidence different oral reading fluency difficulties. *Language, Speech & Hearing Services in Schools (Online), 41*(3), 340-349.

Wong, K., & Nicotera, A. (2007). *Successful Schools and Educational Accountability*. Boston, MA: Pearson.

Woodbridge, M. W., Sumi, W. C., Yu, J., Rouspil, K., Javitz, H. S., Seeley, J. R., & Walker, H. M. (2014). Implementation and sustainability of an evidence-based program: Lessons learned from the PRISM applied to First Step success. *Journal of Emotional and Behavioral Disorders, 22*(2), 95-106.

Xu, Z., Hannaway, J., & D'Souza, S. (2009, June). *Student transience in North Carolina: The effect of school mobility on student outcomes using longitudinal data.* Presentation for the Workshop on the Impact of Mobility and Change on the Lives of Young Children, Schools, and Neighborhoods. Washington DC: The National Academies.

Yeo, S. (2010). Predicting performance on state achievement tests using curriculum-based measurement in reading: A multilevel meta-analysis. *Remedial and Special Education, 31*, (6), 412-422.

Yovanoff, P., Duesbery, L., Alonzo, J., & Tindal, G. (2005). Grade-level invariance of a theoretical causal structure predicting reading comprehension with vocabulary and oral reading fluency. *Educational Measurement, Issues and Practice, 24*(3), 4-12.

Zvoch, K. (2009). Treatment fidelity in multisite evaluation: A multilevel longitudinal examination of provider adherence status and change. *American Journal of Evaluation, 30*(1), 44-61.

Zvoch, K. (2012). How does fidelity of implementation matter? Using multilevel models to detect relationships between participant outcomes and the delivery and receipt of treatment. *American Journal of Evaluation, 33*(4), 547-565.

Zvoch, K. and Stevens, J. (2003). A multilevel, longitudinal analysis of middle school math and language achievement. *Education Policy Analysis Archives, 11* (20).