Title
An Investigation of the Effectiveness of an Orton-Gillingham Based Reading Intervention in Kindergarten and First Grade Using a Fuzzy Regression Discontinuity Design

Author
Patterson, Daniel

Publication Date
2016

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
RIVERSIDE

An Investigation of the Effectiveness of an Orton-Gillingham Based Reading Intervention in Kindergarten and First Grade Using a Fuzzy Regression Discontinuity Design

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Education

by

Daniel Lee Patterson

June 2016

Dissertation Committee:
Dr. Gregory Palardy, Chairperson
Dr. Rollonda O'Connor
Dr. Thomas Smith

The Dissertation of Daniel Lee Patterson:

Committee Chairperson

University of California, Riverside

ABSTRACT OF THE DISSERTATION

An Investigation of the Effectiveness of an Orton-Gillingham Based Reading Intervention in Kindergarten and First Grade Using a Fuzzy Regression Discontinuity Design

by

Daniel Lee Patterson

Doctor of Philosophy, Education
University of California, Riverside, June 2016
Dr. Gregory Palardy, Chairperson

Delays in the development of early literacy skills are associated with a wrath of negative educational outcomes and so addressing such delays is one of the most pressing challenges in education. This study examines the effectiveness of the Orton-Gillingham (OG) Method, a multisensory reading program where instruction utilizes two or more senses simultaneously. Originally developed in the 1930s as a program for dyslexic students, OG has seen continual use since its creation and is endorsed by the American Dyslexic Foundation and the International Dyslexic Foundation. Over the past two decades OG has increasingly been incorporated into general education settings in the primary grades as a reading intervention for struggling readers regardless of whether they have dyslexia. However, there is a dearth of research demonstrating its causal effect as a reading intervention for children with dyslexia or who are experiencing reading delays for other reasons. Two quasi-experimental methods, Regression Discontinuity Design and Nonequivalent Comparison Group Design with propensity scores, are used to test the efficacy of an OG-based, general education reading intervention on a sample of over 700 kindergarten and first grade students who are experiencing reading delays from a large district in California. The Dynamic Indicators of Basic Early Literacy Skills (DIBELS) assessments were used to assign students to the intervention and measure their end-of-year reading outcomes. The results of both analyses revealed no effect for students enrolled in the intervention in either kindergarten or first grade. Within the year that students received the intervention, a small but non-significant gain on end-of-year DIBELS composite scores was found. Long-term outcomes showed that over half of the students in the intervention were still not meeting reading targets by the end of second grade. Moreover, while the treatment effect was found to vary significantly across classrooms and across schools, no available measures classroom or school characteristics where associated with that variation. These findings suggest that certain applications of the OG methodologies may not be effective in general education settings.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION

Introduction

The chapter introduces the problem of the poor rate of literacy in the United States and the need for early identification and remediation for those students at-risk of not developing reading skills in a timely manner. This is followed by promising research that shows most at-risk children respond well to structured, explicit intervention focusing on a well-defined set of early literacy skills. The justification for studying the effectiveness of an early literacy intervention, particularly in an urban district where poor reading rates are even more pronounced, will be presented. An explanation of the present study, which provides a rigorous evaluation of a reading intervention for kindergarten and first grade students at risk of reading failure, will be provided. The intervention is based on the Orton-Gillingham method, a popular instructional method, but one with few rigorous studies providing evidence of its efficacy. The policy implications of early literacy instruction in the present accountability system within California will be discussed. Finally, this chapter provides the purpose of the study, followed by four research questions.

The Importance of Early Literacy Skills

Learning to read is the most important skill acquired by students during the early elementary grades. The acquisition of a well-defined set of foundational early literacy skills is highly predictive of later reading success (Torgesen, 2002). Not only do these early literacy skills provide access to later proficiency in reading, the reading achievement a child acquires in school can have lasting effects throughout life (Kutner, Greenberg, Jin, Boyle, Hsu & Dunleavy, 2007). Reading ability in 3rd grade is predictive of school completion (Lloyd, 1978) and as a consequence it has been argued that early literacy interventions should be a key component of dropout prevention programs (Reschly, 2010). Early reading difficulties are highly associated with many key dropout indicators, such as grade retention (Snow, Burns, and Griffin, 1998) and placement in special education (Donovan & Cross, 2002). The link to literacy and adult well-being has also been shown. In 2003, The National Assessment of Adult Literacy (NAAL) found that adults with low literacy skills vote less, have lower wages, and lower rates of full-time employment (Kutner et al., 2007). Additionally, adults with low literacy skills have more difficulties meeting health-care needs of themselves and their families, and have more trouble with the law (Lonigan & Shanahan, 2009).

Despite the inarguable importance of literacy skills, reading achievement in the United States remains low. Current large-scale assessment results provide evidence of this phenomenon. The National Assessment of Educational Progress (NAEP) has been measuring reading and math achievement since the late 1960s. In the most current year of published results (2013), the rate of students reading proficiently at fourth grade is 34% and 36% at eighth grade (NCES, 2013). Reading at such below-basic levels do not allow for the sufficient completion of schoolwork (Lee, Grigg, & Donahue, 2007) and hinders the educational success of many students.

The Stability of Reading Trajectories

Not only is reading achievement low nationally, initial differences in students' reading ability when entering school often remain throughout schooling. Researchers have routinely reproduced studies that have shown this stable trajectory of reading achievement in students over the course of their education (Catts, Bridges, Little & Tomblin, 2008; Cain & Oakhill, 2011; Francis & Shaywitz, 1996). The troubling consequence of this body of research is that students who struggle in the early grades are more likely to continue struggling through the remainder of their schooling. An early reference to this issue was made by Walberg and Tsai in 1983 who proposed that individuals who have advantageous early educational experiences are able to utilize new educational experiences more efficiently. In terms of reading acquisition, Stanovich (1986) referred to this "rich get richer" phenomenon as the *Matthew Effect*. Simply put, early deficits in reading skills will cause a student to access less text, which further hinders their reading development. Conversely, those students who read efficiently are exposed to more text which further enhances their reading skills. It also allows them to more quickly and at an earlier age begin "reading to learn" rather than, vice-versa, "learning to read." For this reason, students with early reading difficulties tend to lag in other academic areas that depend on reading to learn (Adams, 1990; Reschly, 2010).

A seminal work describing this stable reading growth of students was conducted by Juel in 1988. She tracked the reading achievement of children from first to fourth grade, annually assessing the cohort on phonemic awareness, decoding, word recognition, reading, listening comprehension, reading comprehension and other cognitive measures. Juel found that students' performance in reading in first grade was highly indicative of their achievement in fourth grade, with the top and bottom quartiles in reading achievement consisting of predominantly the same students each year. The bottom quartile in fourth grade contained 21 of the initial 24 students identified as poor readers in first grade. Almost identical patterns were found for good readers, with 26 of the 30 students still in the top quartile in fourth grade. The high proportion of students remaining either poor or good readers throughout early elementary grades, almost 90% in both cases, has been replicated repeatedly in other studies that followed. While the stability of the relative performance of a small cohort of students to each other is not necessarily problematic in itself, the issue is that the vast majority of initially poor readers never met thresholds of reading proficiency within the timeline of the study. At the end of first grade, the good readers had experienced about twice the connected text in basal readers as the poor readers. As a consequence, by fourth grade, the poor readers had not met the same level of decoding skills that the good readers had by the end of second grade. While Juel's sample size was small, later studies reinforced her findings.

More recently, Good, Simmons, and Kame'enui (2009) tracked several cohorts of students from kindergarten through the third grade and found similar proportions of students consistently missing grade level targets on reading skills. They used a battery of assessments on foundational literacy skills called the *Dynamic Indicators of Basic Early Literacy Skills* (DIBELS) to provide on- and off-track indicators of the progression of students' early reading skills. The purpose of the study was to investigate the utility of early assessment indicators for identifying students at risk of later reading problems, in this study measured by poor performance on a high stakes reading assessments in the third grade. The third grade assessment was the Oregon Statewide Assessment, a required assessment given to all students to satisfy the requirements in the No Child Left

Behind Act of 2002. The students were assessed on onset recognition fluency in kindergarten, phoneme segmentation fluency and nonsense word fluency in kindergarten and first grade, and oral reading fluency in first through third grade. The authors set a performance cut-off or *benchmark* on each assessment that if met, would indicate a high likelihood that later benchmarks would be met. They found that more than 90% of students who met an earlier benchmark met the later benchmarks in the corresponding developmental skill. Additionally, they found that students who were identified as needing instructional support met the later benchmark less than 30% of the time. The researchers found that of the 191 students who met the first grade benchmark, 96% also met the proficiency benchmark in third grade. The authors concluded that using benchmarks beginning in first grade identified struggling readers in later grades with high precision. Although the assessments used by Good et al to measure successful reading varied from those used by Juel (1988), the results provided further evidence that a student's performance on pre-reading skills in early elementary is consistent with their reading achievement years later.

When students' reading achievement was tracked for longer periods of time, early reading problems were also found to persist. Using data from the Connecticut Longitudinal Study, a representative sample of 445 children entering public kindergarten in Connecticut in 1983, Francis, Shaywitz, Stuebing, Shaywitz, and Fletcher (1996), investigated two competing growth models of reading achievement, comparing the growth of poor readers with average readers. The first model, the *lag model*, proposed that poor readers have some cognitive skills that develop at slower rates than their peers, which inhibits their reading. This model proposed that the lagging skills will eventually develop, just at a slower rate for some students. If this model is correct, poor readers would eventually catch up to their better performing peers and plateau in skill at a similar level, just at a later time. Consequently, reading intervention would be unnecessary, or at least of only minor importance, as students would eventually gain the requisite reading skills. The second model, the *deficit model*, proposed that poor readers have skills that develop so slowly it leads to lower reading achievement. Without the development of these skills, the students' reading achievement will never catch up to that of their peers, resulting in perpetually poor reading skills. In this model, remediation of missing skills is vital for students to reach adequate levels of reading achievement.

For their study, 403 students were included with reading scores taken annually over nine years from grades 1-9. Students reading achievement was assessed using the Rasch-scaled reading-cluster score of the Woodcock-Johnson Psychoeducational Test Battery. The Rasch scale used in the assessment provides a grade level equivalent result. For example, a score of 500 indicates the average achievement of a beginning fifth grader. This sample consisted of three groups. The largest group (n = 301), were students that exhibited no reading deficiency in first grade. The students in the two other groups had reading deficiencies, but were identified in different ways. The first deficient group was identified using a discrepancy model, which is routinely used to screen students for special education services. These students had reading achievement at least 1.5 standard deviations below what would be anticipated by their IQ, which is a common definition of a reading disability. The second deficient group had reading achievement consistent with IQ, but were in the bottom 25% of the sample.

The researchers found that a quadratic to plateau model fit all students' reading growth over time. The growth was nonlinear, growing quickly in the early years then slowing as the children got older, as seen in Figure 1. Furthermore, analysis of the trajectories of the three groups supported the lag model. The two deficient groups, by definition, started with initial scores well below the average reading group. These two groups exhibited higher initial growth in reading achievement as measured by the Woodcock-Johnson. However, this growth slowed and all groups plateaued around the same time, with students reaching their maximum reading achievement around the age of 15. The deficient groups never reached the level of reading of the average group. The deficient group plateaued on average at the level of reading that the proficient groups had reached by age 9. In a study of over 600 children from 2nd through 10th grade Catts, Bridges, Little and Tomblin (2008) also found that the deficit model fit the reading growth of the students in their study.

Reading Intervention Studies

The consistent finding that trajectories of reading growth are stable, with the consequence that those who begin as poor readers rarely become good readers, led many researchers to investigate the effects of early identification and remediation of students at risk of reading failure. Promisingly, students identified as at-risk of reading failure in the early grades have responded to reading intervention efforts with high rates of success (Gerston, Compton, Connor, Dimino, Santoro, Linan-Thompson & Tilly, 2009; Denton, 2012). While previous longitudinal studies showed that poor readers most often remain poor readers with only their typical instruction, several longitudinal intervention studies have shown that with proper intervention strategies these trajectories can be positively altered for the majority of poor readers (Wanzek & Vaughn, 2007). Lyon, Fletcher, Shaywitz, Shaywitz, Torgesen, Wood, Schulte and Olson, R (2001), note:

> “we have learned that for 85 to 90 percent of poor readers, prevention and early intervention programs that combine instruction in phonemic awareness, phonics, spelling, reading fluency, and reading comprehension strategies provided by well trained teachers can increase reading skills to average reading levels (p. 1)”

Additionally, evidence supports that the earlier the students are identified and remediated the more successful the results are (Wanzek & Vaughn, 2007).

While not as high as Lyon’s estimates, Torgesen (2000) showed that the majority of students placed in early reading interventions improved their reading achievement. Five intervention studies were reviewed that had been conducted in kindergarten and first grade. Torgesen’s analysis involved students in the bottom 12-18 percent of kindergarten and first grade. The five studies he reviewed were chosen because they were well-designed, implemented with fidelity, and followed instructional programs based on early literacy research. All five studies reviewed resulted in a reduction in the number of children at risk of reading failure. The estimates showed that early intervention can reduce the number of children who develop reading difficulties by up to 70 percent.

Vellutino, Scanlon, Small, & Fanuele (2006) sampled 1,373 students entering kindergarten, identifying those at risk or not at risk for early reading difficulties on tasks including letter-name knowledge and phonological awareness. Half of the identified children received an intervention designed by the researchers two to three times a week while the other half received the intervention offered at their school. At the beginning of first grade, all of the identified students were assessed again for reading difficulties, with little difference in outcomes based on the intervention used. Students that were below the fiftieth percentile continued to receive intervention. As in kindergarten, half of the students received instruction designed by the researchers and the other half received what was offered by the school. Of the students that received either kindergarten only or both kindergarten and first grade interventions (n = 117), 84% were no longer at risk by the end of third grade. The majority of this group received only a single year of remediation in kindergarten.

The age that students receive intervention has been shown to influence their rate of success, with earlier interventions having greater success rates than those starting later. In particular, reading interventions conducted in kindergarten and first grade have been shown to be more effective than interventions at later grades. Wanzek and Vaughn (2007) conducted a meta-analysis of 18 early reading interventions conducted in kindergarten through third grade between 1995 and 2005. The authors found that interventions provided before second grade produced the largest effect sizes. The authors noted that "the difficulties students face in reading as they enter second or third grade are more complex than in first grade, making significant gains more difficult to achieve" (p. 557). They also note that children who are not identified as having reading difficulties until the third grade are most likely those with the most severe difficulties, making remediation more challenging.

More recently, Simmons, Coyne, Kwok, McDonagh, Harn, and Kame'enui (2008) showed that students responded early and successfully to early intervention and that kindergarten is a critical response period. They tracked 41 students who were identified as poor readers in kindergarten. Students were identified as in need of intervention by falling below the 30th percentile on screening assessments, based on nationally normed results. Because of the developmental nature of reading, different measures were used at different times. The initial population was identified by naming less than five letter names within a minute at the beginning of kindergarten. Later assessments included measures from the DIBELS assessments and reading components from the Woodcock-Johnson examination. Similarly, O'Connor, Fulmer, Harty, and Bell (2005) found that students who participated in ongoing small-group or individual instruction showed decreases in reading disability at the end of third grade.

All this indicates that early assessment and interventions are critical. Although the previous studies used different indicators of risk and different instructional methods for the interventions, the findings are promising. Further, the window found within kindergarten and first grade seems to be crucial to begin interventions for students that begin school at risk of reading failure. If the opportunity is missed, the consequences are potentially lasting and severe. Unfortunately, many schools still do not have structured interventions in place for children at risk of reading failure in the early grades. For those that are beginning to implement programs to address this need, using programs aligned to research and evaluating their efficacy is critical.

Purpose of the Study

Given the importance of early literacy skills and the higher rates of at-risk students in urban districts, the need for high quality reading interventions is paramount. A large urban school district recently underwent a structured implementation of a kindergarten and first grade early literacy intervention based on the Orton-Gillingham Method. The Orton-Gillingham method is a multi-sensory reading program that was originally developed for dyslexic students over fifty years ago and has more recently been adapted for general education classrooms. While Orton-Gillingham is a popular method, there is little well-designed research showing its effectiveness. The district has piloted the program at a small number of schools with the plan to scale the intervention district wide. Expanding a program to such a wide scale with no evaluation of its efficacy is problematic. This study will use a regression discontinuity (RD) design that will meet the standards of the What Works Clearinghouse to investigate the effectiveness of this intervention on both the short term and long term reading outcomes of the students that participate in the intervention.

Policy Implications

The preceding evidence highlights the critical importance of schools properly identifying students who enter school at-risk of reading failure and intervene with well-structured programs based on the research on early literacy skills. Investing in the development of a robust early reading intervention is likely an imperative for large urban districts, who service large populations of at-risk youth. Although schools have limited resources to direct to a wide educational interventions, recent changes in California finance policy have provided school district with greater flexibility in how they use state funds. In 2013, California did away with its decades old funding formula based on horizontal equity and replaced it with a weighted student formula known as the *Local Control Funding Formula*. Districts now have greater flexibility in how they direct money due to the elimination of dozens of categorical programs that carried with them strict allocation rules. To provide transparency with how they spend money, districts must write a *Local Control Accountability Plan*, or LCAP that explicitly shows the amount of funds used to support different programs. Every district must identify gaps in educational outcomes based on data and local input from the community and direct funds at increasing achievement in those areas.

This change opens up the accountability system to grade levels that were not previously focused upon. Previously, neither state nor federal accountability systems began measuring student achievement until the second or third grade. Under the new state accountability system, any identified need can be addressed with goals, programs and funding. The large urban district in this study has incorporated a ‘reading by third grade’ initiative into its LCAP. This goal is primarily being met through the implementation of the Orton-Gillingham based intervention. Children identified as at-risk of reading failure in kindergarten and first grade are provided additional structured lessons that target early literacy skills. The district has invested a substantial amount of funds in materials and teacher training to implement the program. Ensuring an early literacy program is effective, given the importance of reading skills and the limited resources districts have to address problems, is vital to the long term success of students.

Research Questions

The three research questions are as follows:

1. Is the Orton-Gillingham based reading intervention an effective reading intervention for kindergarten and first grade? Does the program produce a significant effect size in reading achievement for students who participate?
2. Does the efficacy of the program differ for kindergarten and first grade?
3. Do the effects of the intervention last? Do students at the end of 2nd grade show adequate reading skills?

CHAPTER 2 LITERATURE REVIEW

Introduction

This chapter reviews the research related to an Orton-Gillingham based reading intervention for students in kindergarten and first grade. The intervention, henceforth referred to as the OG Institute, was implemented over the course of several years in a large urban school district in California, with a small number of schools implementing it each year. The first section of this chapter outlines the Orton-Gillingham (OG) method which has been in use as a reading intervention nationally for decades and is still a popular method for instructing students with reading difficulties. The instructional components of the program will be examined. Specifically, the multisensory instructional method and the early literacy skills taught in the program will be reviewed. The research on the progression of skills children should acquire in kindergarten and first grade as they learn to read will be presented. The progression of these skills, the typical timeline that they develop, and the common barriers children encounter as they learn to read will be discussed. The lack of rigorous research on the effectiveness of the Orton-Gillingham method will be addressed in the second section, clearly showing that a rigorous evaluation of the program for the target population is needed.

The Orton-Gillingham Method

History

Few instructional programs have had the continuous widespread use and influence as seen by the Orton-Gillingham (OG) Method. Instruction based on OG has been in existence since the 1930s and has been steadily implemented in schools for over half a century, primarily as a reading intervention for dyslexic students (Ritchey & Goeke, 2006). More recently, the instructional program has transmuted from its original conception as a special education reading intervention into general education classrooms. The method is rooted in the work of Samuel Orton, a physician who began studying ‘word blindness’ in patients in the mid-1920s and became intrigued with the reading problems encountered by bright children (Henry, 1999). Orton’s research on dyslexia spanned four decades and influenced much of the early work in the field. Orton theorized that dyslexia, which at the time was characterized by twisted symbols, or strephosymbolia, was caused by a lack of dominance in the brain hemispheres resulting in mirror images. Orton’s theory on the causes of dyslexia were not supported by later research, however. The current consensus is that dyslexia is caused by phonological processing problems, not by brain dominance (Ritchey & Goeke, 2006). Despite the shortcomings of his theories on dyslexia, the instructional method he developed to aid dyslexic students in their development of reading and writing is still widely used.

In the early 1930s, Orton requested Anna Gillingham to organize his neurological hypothesis into an instructional program (Henry, 1998). Gillingham enlisted the help of her colleague, Bessie Stillman and together the first edition of their work, *Remedial Training for Children with Specific Disability in Reading, Spelling, and Penmanship*, was released in 1936. Later editions were regularly released and it is currently in its eighth edition (1997). The Orton-Gillingham Method is “a multisensory explicit phonics method with emphasis on visual and auditory feedback for sounds and the tactile-kinesthetic input of letter formation” (Alexander & Slinger-Constant, p. 751). In the program, children are taught the relationship between letters and sounds through the use of activities that purposely use several senses simultaneously, such as saying a letter out loud while tracing the letter on a plate filled with sand to increase the tactile stimulation. The use of two or more senses at once is theorized by the supporters of multisensory instruction to reinforce learning (Gillingham & Stillman, 1997).

Adaptations of OG Instruction

Over the decades following its first release, the OG method grew in popularity and inspired many adaptations based on the instructional methods outlined in the manual. Rose and Zirkel (2007) noted that the past 30 years have produced more than 15 commercial programs and several private schools for students with disabilities centered on the OG methodology. Currently, the International Dyslexia Association lists several multisensory approaches to language instruction. Though not exhaustive, the list includes: The Orton-Gillingham Approach, Alphabetic Phonics, The Association Method, Language!, The Lexia-Herman Method, Lindamood-Bell, Project Read, Slingerland, Sonday System, Sounds in Syllables, Spalding Method, Starting Over, Wilson Fundations & Wilson Reading. Additionally, the term Multisensory Language Instruction is generally applied to instruction based on the Orton-Gillingham process. Despite these variations, OG and its adaptations share the common thread of a phonics based reading curriculum delivered through multisensory instruction (Ritchey & Goeke, 2006). The Orton-Gillingham method has also been used on a wide range of grade levels, from early elementary to college. While a review of the research on every adaptation of OG would yield similarly sparse results, this review will be limited to OG based programs targeting the population addressed in this study - general education students in kindergarten and first grade.

Lack of Methodologically Rigorous Research

Despite the support of Orton-Gillingham based instructional methods by large organizations such as the International Dyslexia Association and the American Dyslexia Association, there are surprisingly few rigorous studies demonstrating its efficacy. This is not to say that studies have found OG to be ineffective, but that the studies investigating its instructional merit have not been well designed. It appears that the widespread use of OG instruction has been driven by anecdotal evidence and personal experience (Ritchey & Goeke, 2006). While studies on OG have been conducted for decades, they have predominantly been flawed and there exist "few methodologically sound studies in peer reviewed journals to validate its efficacy" (Alexander & Slinger-Constant, 2004, p. 751). In a review of OG based litigation by parents specifically requesting OG as their intervention of choice, Rose and Zirkel (2007) claim there is a "gap that persists between professional advocacy for OG and the need for scientifically based and peer-reviewed research (p. 174)." This sentiment was similarly echoed by Alexander and Slinger-Constant (2004) who stated that "...future research using methodologically sound designs is badly needed (p. 751)". More recently, the lack of research on OG was further reinforced by a series of What Works Clearinghouse Intervention (WWC) Reports, a division of the United States Department of Education, released between 2010 and 2012. Of the over 150 OG and OG based instructional studies reviewed, only two met the organization's minimum requirements for methodologically sound studies. A current review of research on OG, which follows, will show that the gap in scientific evidence on the efficacy of OG still exists and that a causal link between OG based instruction and increased reading achievement has yet to be established.

Instructional Components of OG

Multisensory Instruction

The most notable difference between OG and other phonics programs is its purposeful inclusion of multisensory instruction (MSI). The teacher is directed "to assist children in making numerous visual, auditory, and kinesthetic-tactile linkages, as portrayed by the 'language triangle' (Henry, 1998). The language triangle (Figure 2) is an organizational device that represents the connections between three different modes of learning; its vertices denoting the *Auditory* (A), *Visual* (V), and *Tactile-Kinesthetic* (T) modes of learning. In OG, multisensory tasks are referred to as *associations,* which intentionally utilize two or more senses during learning. Early researchers such as Montessori (1912), and Fernald and Keller (1921), and Strauss and Lehtinen (1947) theorized that using multiple senses simultaneously provides a richer opportunity for learning that reinforces the lessons in memory. When utilizing this method, "every phonogram (representation of a sound) is presented through each association (visual, auditory, and kinesthetic), and each association is linked and presented simultaneously. The individual pathway makes an imprint on the brain and thus strengthens the learning process" (Gillingham & Stillman, 1997, p. 30). For example, "... if the student's visual reinforcement is looking at a letter, the auditory reinforcement is to listen to and hear the sound and identify it with its symbol. The kinesthetic reinforcement stems from the child's feeling the articulatory muscle movement associated with saying the phonemes, as well as writing the letter on a roughened surface and feeling the associated sensations" (Birsch, 2005, p. 25). The complete set of the multisensory lessons from the OG institute that reference these associations is in the appendix.

Multisensory instruction has a strong constituency of practitioners and is endorsed by the American Dyslexia Society. Additionally, the International Dyslexia Association, a professional organization devoted to dyslexia and related reading difficulties, has endorsed multisensory methods used in combination with structured language teaching (see www.interdys.org/factsheets.htm), claiming that multisensory programs are especially appropriate as small group, Tier II or III interventions within a Response to Intervention (RtI) framework. However, scientific support of MSI is more theoretical than empirical, with "the appeal of multisensory instruction endures even though it has been poorly defined and is not well validated in existing intervention studies" (Birsh, 2005, p. 33).

Despite the current support of multisensory instruction by organizations and practitioners, sparse empirical evidence exists for MSI (Uhry & Clark , 1995; Birsch, 2005). Little research has isolated the portion of a reading intervention's effectiveness that can solely be based on the MSI component of the instruction. The merits of including MSI in a reading program are still debated (Spear-Swerling, 2013). In studies that have investigated multisensory techniques, none have isolated the influence of the multisensory methods (Uhry & Clark, 2005; Fletcher, Lyon, Fuchs, & Barnes, 2006). For example, Wise, Ring, and Olson (1999) did not find that the multisensory articulatory component of the Lindamood program was necessary when they compared it to the results that did not include it.

However, as noted by Spear-Swerling (2013) supporters of multisensory techniques argue it provides focus on letters and print and the use of a multisensory tracing technique ensures that students are looking at the letter at the same time that they say its sound. Additionally, the intense, systematic approach of these programs and their explicit attention to the structure of language may be what is so beneficial to struggling readers. Although benefits of multisensory instruction are not clear, the research on the core components of early literacy instruction is more solid. In particular, the use of a structured synthetic phonics based program, as is used in most multisensory programs, has been supported by research over the last few decades.

Research on Early Literacy Instruction

The National Reading Panel (2000) reviewed five core components of early literacy: phonemic awareness, phonics, fluency, vocabulary and comprehension. Each of these components was found to be important to later reading achievement. Based on this review and the recommendations and of others, such as Adams (1990) and Griffin, Burns and Snow (1998), the No Child Left Behind Act of 2002, through its Reading First Initiative, formally defined that a quality reading program should have these five components (Title I, Part B, Subpart 1) stating that public schools are required to "establish reading programs for students in kindergarten through grade 3 that are based on scientifically based reading research, to ensure that every student can read at grade level or above not later than the end of grade 3*" (§1202[c][7][A]).*

Both general education core curriculum and interventions for students at risk of reading failure should be designed with this progression in mind. According to O'Connor, Fulmer, Harty, and Bell (2005):

> In essence, the rationale behind most early intervention efforts is to encourage at-risk children to learn the earlier understandings about sounds in words and features of print within the windows in time that these understandings develop for typically achieving children. In this view, if by the end of kindergarten all children could blend and segment spoken words (phonemic awareness) and link these speech sounds to letter sounds (the alphabetic principle), they would be prepared to learn to decode and recognize words in first grade. Likewise, if children learned to decode and recognize words in first grade, alongside maintaining adequate vocabulary growth, they would be prepared for building reading fluency and more complex comprehension in later grades (p. 440)

This review will outline the core skills noted above that should be developed in kindergarten and first grade, namely phonemic awareness, the alphabetic principle, and learning to decode with phonics. Their incorporation into the ten step lessons plan used in the OG institute will be given. The full set of lessons used in the OG institute is found in the appendix. In these lessons typically six students meet for twenty minutes a day with their classroom teacher. The program is structured so that roughly two lessons are covered each day and that within the five day week all ten lessons are completed. The lessons repeat following the next topic (letter or blend) in the sequencing chart, also provided in the appendix. The first lesson in the kindergarten and first grade OG Institute addresses phonemic awareness skills. Several other lessons address phonemic awareness skills but include the use of letters. The research on the benefits of including letters in phonemic awareness training will be reviewed later in the chapter. The research on phonemic awareness instruction will be addressed first.

Phonemic Awareness

English is an alphabetic language, where graphemes (written letters) translate into phonemes (spoken sounds). Reading requires an understanding of this principle, but before students can begin to decode printed letters on a page into spoken words, they must first understand that spoken words are broken into smaller units of sound. This realization does not come naturally to young children solely through their experience with oral language. While learning to talk occurs naturally for most students, learning to read does not, and is difficult to the point that it has been called an “unnatural act” (Gough, 1980). One factor impeding understanding is that the individual sounds of words are not easily distinguishable in speech. It was Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967) who, using spectrographic analysis, established that phonemes are coarticulated in the speech stream. The result is that we hear words such as ‘bag’ not as three separate phonemes or sounds, but as one single acoustic unit. As a consequence, some children require explicit instruction in phonemic awareness to break apart what they hear into its constituent sounds and the letters that represent them, which is necessary for them to translate between the letters in words and the sounds they hear.

Phonological awareness is an oral language skill that manifests itself in the ability to notice, to think about, or manipulate the individual sounds in a word (Anthony & Francis, 2005). It is seen in such skills as syllabification, rhyming, and onset-rime identification that recognize the smaller sound units within words. *Phoneme awareness* is a subset of phonological awareness that represents an understanding of the smallest units of sound in words that have meaning, called phonemes. The English language comprises 44 phonemes, each of which can be represented by one or more letters. Although each phoneme can be represented by letters, phonemic awareness itself does not require understanding letters; rather, it is sound play (O’Connor, 2014). While phonological awareness as a whole is important to reading, having phonemic awareness is critical for children to be successful readers (Schuele & Boudreau 2008).

Phonemic awareness is displayed through manipulating the phoneme level of words. This can be displayed through various skills such as identity, blending, segmenting, deletion, and substitution; each of which can be taught and assessed in many ways. The two skills most fundamental to reading are blending and segmenting (Schuele & Boudreau, 2008). Blending is the ability to take isolated phonemes and elongate their sounds, slowly merging them into one another to form a word. An example would be forming the word ‘sun’ from individual phonemes: s – u – n → sun. Working in an opposite direction of blending is the phoneme awareness skill of segmentation, which is the ability to take a spoken word and break it into its component phonemes. An example would be understanding the word ‘that’ can be broken into three separate phonemes: th – a – t, and identifying or saying each individual sound in isolation. Blending relates to reading unfamiliar words, while segmenting occurs with spoken words and is more related to spelling. Children can exhibit phonemic awareness through phoneme deletion and substitution as well. An example of deletion would be asking a child to say the word ‘cat’ without the “t” sound. Lastly, substitution is a skill that can be demonstrated by a child replacing a single phoneme in a word with a new phoneme. An example would be saying the word “cat” with an “m” instead of a “c.”

Correlational Research on Phonemic Awareness. In an effort to identify which measurable traits in pre-reading children led to reading success, researchers began investigating which pre-reading skills measured in the primary grades were predictive of later reading success. These early efforts, many of which were conducted in the 1980s, were motivated by the belief that early remediation would reduce the number of children who become diagnosed with reading disabilities in later grades. This work was correlational in nature, measuring many traits of pre-readers in kindergarten and first grade to find which skills were most predictive of later reading achievement. Researchers repeatedly found that phonemic awareness was consistently the most predictive of later reading success (Share, Anthony, Jorm, Maclean & Matthews, 1984; Lundberg, Frost & Petersen, 1998).

One early study conducted by Share et al in 1984 showed phonemic awareness, compared to numerous other skills in young children, was the most highly correlated to later reading achievement. The researchers measured thirty-nine indicators, including pre-reading skills, oral language ability, motor skills, family background characteristics, and social behavior on a group of 584 Australian students entering school for the first time. Reading achievement was assessed at the end of the students' first and second years of schooling. Phonemic awareness was measured by two tests. The first test required the student to say the initial phoneme of the word and then the remaining part of the word with the initial phoneme removed. The second test required the students to segment a word into its initial, medial, and final phonemes. Using simple correlations, the researchers found that out of the 39 measures, phonemic awareness was the most predictive of reading achievement at the end of their first and second year of school, being 0.66 and 0.62 respectively. Close in correlation was letter identification, which has also been seen to be highly indicative of early reading skills. The authors concluded:

> The fact that the phoneme-segmentation task was as predictive of reading achievement as the letter-names task is an important finding that represents strong support for theorists who have argued for an important role of phonemic awareness in learning to read. (p. 1313)

Additional correlational support for the importance of phonemic awareness was added by Tunmer and Nesdale in 1985. Their study included 63 Australian first grade students. The students were administered tests on segmentation, decoding, and reading comprehension. The study found segmentation to be strongly correlated to reading comprehension. The decoding tasks were the most correlated to reading comprehension, the correlation being .89 and .83 of real and pseudowords respectively. Further strengthening the understanding that phonemic awareness was critical for reading, the researchers found important relationships between segmentation, decoding, and comprehension. In their sample, of the 22 students who passed the decoding test, all had passed the segmentation test as well. Additionally, of the 24 students who failed the decoding test, none passed the segmentation test. This led the authors to propose that segmentation was a necessary for students to develop later decoding skills.

Juel (1988) conducted another correlational study measuring several traits in 54 low SES students, tracking them from first through fourth grade. Throughout this period, students were assessed in phonemic awareness, decoding, word recognition, reading, listening comprehension, reading comprehension and other measures. Juel found that the students who were poor readers in fourth grade had entered first grade with little phonemic awareness. As a consequence, the poor readers had been exposed to about half the connected text as good readers by the end of first grade, resulting in poor readers remaining poor readers. Of the 24 students in the bottom quartile in reading in the first grade, all but three were still poor readers by the end of fourth grade. The phonemic awareness skills that good readers exhibited in first grade were not reached by the poor readers until the end of third grade. Additionally, the decoding skill that poor readers developed by the end of fourth grade was not at the same level of good readers at the end of first grade. The work of Juel and Tunmer and Nesdale provided evidence that phonemic awareness was highly predictive of later reading success and that it was a necessary, but not sufficient, skill for later reading.

Causal Research on Phonemic Awareness. Although these and other studies (Felton, 1992; Perfetti, Beck, Bell, & Hughes, 1987) found that children with better phonemic awareness skills eventually became better readers, they did not prove that improvement in phonemic awareness would improve later reading ability. To see whether increasing a child's phonemic awareness skills would improve his or her success in later reading, researchers began manipulating children's phonemic awareness skills in controlled experiments to study the effect on reading. Early efforts began decades ago through the work of Elkonin in 1965. Using *Elkonin Boxes*, which are rectangular boxes divided into components in which students place individual phonemes, he provided phonemic awareness training to students. Although his studies did not include data on the results, his work was influential, and Elkonin Boxes are still used today in instruction. The OG institute uses Elkonin Boxes in several lessons, including its first lesson on phonemic awareness. The complete lesson can be found in the appendix. Fortunately, the link between phonemic awareness instruction and increased reading success was made by many researchers in the later decades.

Bradley and Bryant (1983) were among the first researchers to show the causal link between phonemic awareness and later reading ability. Their study consisted of two components. The first was the measurement of phonemic awareness skills in 403 students who were entering school for the first time. Reading ability was subsequently measured at the end of their first and second year of school. The researchers used *sound* *categorization*, a form of phoneme identity, as their measure of phonemic awareness. This was measured using an odd man out strategy where children were given three or four words in which all but one shared a common phoneme and they were asked to detect the odd word. Similar to other studies, phoneme awareness was highly predictive of later reading achievement.

The second part of the study consisted of training half of the children on sound categorization over the next two years, leaving the other half as a control. The training group was split into two groups, both receiving 40 lessons in sound categorization over two years. This training consisted of giving students sets of words with either the same initial, medial, or final phoneme, and teaching them to categorize them based on these similarities. Half of the training group had additional instruction that included lessons on letters and their associated phonemes. The control group was also split into two groups; the first half received training in categorization, but it was conceptual in nature, where the students grouped words into categories such as animals and shapes. The second half of the control group received no additional or special instruction.

At the end of four years, as well as periodically throughout the study, the researchers measured all students on spelling and reading achievement as well as on IQ. As in previous studies, the authors found that children who began the study with higher phonemic awareness had higher reading achievement four years later, providing further correlational support linking phonemic awareness to future reading success. Of greater significance were the results of the training experiment which showed that phonemic awareness instruction led to improved reading outcomes. Bradley and Bryant (1983) found both of the groups that received sound categorization instruction had higher achievement in both reading and spelling. The group that received additional support with letters had even higher gains in both reading and spelling.

Other studies reinforced that instruction in phonemic awareness would lead to increased reading achievement. In an experiment with 42 kindergarten and first grade children, Cunningham (1990) divided the students into three groups of 14. Two groups received instruction in phonemic awareness while the third served as a control. Both experimental groups received *skill and drill* lessons on segmentation and blending. One group received additional instruction on the *metalevel* of phonemic awareness, explicitly emphasizing its value and application in reading. Both experimental groups outperformed the control group in reading ability, and the group that received additional metalevel training had even larger gains. This provided insight that providing students with the purpose of the skills they were learning provided further benefits.

Fox and Routh (1984) had similar results with their three-group training model. The three groups of roughly 10 kindergarten students who were not proficient in segmenting were given either segmenting, segmenting and blending, or no additional instruction. The group given segmenting and blending training outperformed the students who could already segment, and better than the control group. Additional studies provided evidence that teaching students to blend and segment words improved later reading ability (Lundberg, Frost & Peterson, 1988; O’Connor, Jenkins & Slocum, 1995; Torgesen, Morgan, & Davis, 1992).

Phonemic Awareness Interventions within the Classroom. While the success of these controlled experiments provided evidence that phonemic awareness was a good leverage point for early literacy instruction, there were some limitations to their real world application. First, children were taught by the well-trained researchers outside of the regular classroom. Second, the trainings were provided in one-to-one or small, homogenous groups. Most schools do not have the resources to reproduce such intensive trainings in the classroom. Evidence that phonemic awareness training could be provided by classroom teachers and paraprofessionals in classroom settings was provided by several later studies.

O'Connor et al. (1996) provided whole class and center-based training to students at high-risk urban schools. In this study, teachers delivered instruction to intact groups of students within existing classrooms. Group size was determined by the existing class size. Assistance from other adults was limited to what was normally available. The intervention lasted for two-thirds of the school year. Children in the study were enrolled in one of three types of kindergarten: general, self-contained, and transitional. The general kindergarten contained a few children with mild disabilities, the self-contained kindergarten contained children with mild disabilities, and the transition kindergarten was for students repeating kindergarten. They found that across student types, children who received the training performed better on phoneme-level manipulations and reading and writing outcomes than untreated children.

Blachman, Tangel, Ball, Black, and McGraw (1999) had similar findings in their study of classroom instruction delivered by classroom teachers of kindergarten students.

They provided intervention training to teachers and their teaching assistants to be delivered in small groups in four, low income, inner-city schools in upstate New York. There were 84 treatment and 75 control students. These groups did not differ on age, sex, race, socioeconomic status, oral vocabulary, phoneme segmentation, letter name and letter sound knowledge, or reading at the beginning of the study. The students in the treatment met four times per week for 15 to 20 minutes per day completing 41 lessons, including segmentation activities and letter sound practice, over 11 weeks. At the end of the 11-week intervention, the treatment children significantly outperformed the control group on almost all measures.

What are the most important Phonemic Awareness skills to teach? Research has conclusively found that a student's phonemic awareness is predictive of their later reading success. It has also shown that children with little phonemic awareness entering schools can benefit from classroom activities that target those skills. Later studies clarified the practical application of this knowledge for classroom instruction. Findings included which phonemic awareness skills are necessary, in which order to teach them, and what level of proficiency students need to attain to become good readers.

Slocum, O'Connor and Jenkins (1993) sought to determine whether blending or segmenting was easier for students to learn. Using only spoken words, they taught half of a group of students blending first, while the other half learned segmenting first. After students mastered the first skill, they taught them the other. They found no differences in the time it took for students to learn each skill or transfer from one skill to the other.

O'Connor, Jenkins and Slocum (1995) taught two groups of students different phonemic awareness skills. One group received instruction on blending and segmenting. The second group learned rhyme and deletion in addition to receiving the same blending and segmenting instruction. They found that the students who received instruction in only blending and segmenting learned to rhyme as well as those that were explicitly taught it. Additionally, they found it difficult to teach deletion to pre-readers. Both groups of students outperformed the control group in reading. The authors concluded that instruction in phonemic awareness should include at least blending and segmenting, along with a few letters. Yeh and Connell (2008) taught three groups of four- and five-year old students different combinations of phonemic awareness skills. One group learned to blend and segment with a few letters, another to recognize and produce rhymes, and the third learned no phonemic awareness skills, being limited to instruction on the meanings of new words. The students who learned blending and segmenting also learned how to rhyme without being explicitly taught, which was the only group to acquire a phonemic awareness skill outside the method of delivery. The researchers concluded that certain phonemic awareness skills, such as blending and segmenting, provide transfer to other phonemic skills.

As summarized by O'Connor (2005), phonemic awareness research has provided several insights. The first is that phonemic awareness is causally connected to reading achievement. That is, students must first have phonemic awareness before they can learn to read. The second is that students entering school without phonemic awareness can be instructed in it and make good gains toward the path to reading. Lastly, phonemic awareness can be typically be taught in under 10 hours of instruction using small doses of 10-20 minutes. The OG Institute instruction on phonemic awareness teaches blending and segmenting and aligns with the preceding research on phonemic awareness. Five of the ten lessons in the Orton-Gillingham address phonemic awareness related skills.

While the phonemic awareness is critical for beginning readers, it is not sufficient (Tunmer & Nesdale, 1985). A finding that many of the previous researchers converged upon in their research on phonemic awareness instruction, which was later expanded upon, is that when letters are added to phonemic awareness instruction, children experience even greater gains.

The Alphabetic Principle.

In addition to phonemic awareness, letter knowledge is an important indicator of learning to read (Hulme, Bowyer-Crane,Carroll, Duff, and Snowling, 2012). Evidence of this relationship has been empirically supported by many studies, including Share et al. (1984) who found that knowledge of letters was highly correlated to later reading, closely following the correlation of phonemic awareness. Although it is established that letter knowledge is an independent predictor of word reading, it has been argued that letter knowledge is a better indicator of parental support or of visual-verbal learning ability (Foulin, 2005) and it does not imply that increased letter knowledge causes improved reading skill (Hulme et al., 2012). However, in phonemic awareness training studies, when lessons on letters-sound correspondences were added to the instruction, results improved (Bradley & Bryant, 1983; Bus & van IJzendoorn, 1999; Byrne & Fielding-Barnsely, 1991; O'Connor & Jenkins, 1995; Rack,Snowling & Olsen, 1992; Torgesen, Morgan, & Davis, 1992). It is only through this explicit instruction in sound-letter correspondences that children gain the most fundamental early literacy skill - the alphabetic principle, which is a “systematic and predictable relationship between written letters and spoken sound” (Bursuck & Damer, 2007). When students have this knowledge they begin to understand that written letters map to spoken sounds; they can then begin to unlock written text. Orton had similar insight, although his was decades before the empirical support that follows. In 1925 he noted that “the logical training for these children would be that of extremely thorough repetitive drill on the fundamentals of phonic association with letter forms both visually presented and reproduced in writing, until the correct associations were built up” (Orton, p. 614).

The OG Institute teaches the alphabetic principle by teaching students sound and letter combinations with attention to phonemic awareness. It begins with high frequency letters that can be used in a variety of short words, giving students many opportunities to practice. It is within these lessons that many multisensory associations are used, such as writing letters on roughened surfaces or sand, and tapping out letters. In the 10 step lesson plan, lessons two through eight address the alphabetic principle before advancing into more advanced phonics topics in first grade. The following section reviews the literature that supports these activities.

Early empirical evidence supporting this idea came from Bradley and Bryant (1983). They found children in both their phonemic awareness training groups made better gains in reading than those students who did not receive phonemic awareness instruction, but the phonemic awareness group that included letters in the training had the highest performance of all groups. While the work of Bradley and Bryant (1983) showed that phonemic awareness instruction that included explicit instruction in the relationship between letters and their sounds provided better reading outcomes than phonemic awareness training alone, it did not show whether letter training alone provided later benefits to reading. To address this question Ball and Blachman (1991) included an instructional group whose focus was on letter knowledge without any phonemic awareness training. In their study 89 kindergarten students from six classrooms in three schools were separated into three groups, a control group and two groups that differed in instructional approach. Both treatment groups were instructed four times a week for 20 minutes over the course of seven weeks. Students were given pretests about halfway through their kindergarten year on oral vocabulary, word identification, segmentation, and letter-name and letter-sound knowledge. ANCOVA analysis showed that these students were similar on all pretest measures. One treatment group was trained on segmenting and correspondences between letter names and letter sounds. The other treatment group was trained only in letter names and letter sounds. Only the treatment group that included phonemic awareness training with letter sound instruction significantly improved early reading skills and spelling skills. The authors found that increased letter-sound knowledge, by itself, does not improve segmentation, spelling, and reading skills.

While the results of the Ball and Blachman (1991) study illustrated the importance of the combination of phonemic awareness training with instruction on letter-sound correspondences, it used trained teachers or clinicians brought to the school by researchers, and the instruction was done outside the classroom. To show that the training would have similar effects within a normal classroom, four years later Blachman, Ball, Black, and Tangel (1994) conducted a similar study using teachers trained to provide the instruction within the classroom. The study was conducted in a low income, inner city school that included 84 treatment and 75 control students. The researchers trained kindergarten teachers and their assistants to provide instruction to small groups within classrooms during the regular school day. Lessons similar in design from Blachman et al. (1994) were used for instruction. The results of the instruction were equivalent to those produced by the trained clinicians outside of the classroom, providing evidence that instruction in the alphabetic principle is critical to students' success in reading and that it can be provided successfully within the classroom environment using only the resources of the school. This suggests that the Orton Gillingham Method could produce positive outcomes since it aligns well to this research. Several other studies in the 1990s also found that teaching phonemic awareness, along with some letter and sound correspondences, greatly improved children's reading achievement (Byrne & Fielding-Barnsely, 1991; ; O'Connor & Jenkins, 1995; Rack, Snowling & Olsen, 1992; Torgesen, Morgan, & Davis, 1992).

More recently Ryder, Tunmer, and Greaney (2008) provided explicit instruction in phonemic awareness and the alphabetic principle as an intervention to at-risk students within a whole-language classroom. Twenty four six- and seven-year olds were provided 24 weeks of highly sequenced lessons by a teaching aid over the course of the intervention. The treatment group significantly outperformed the control on phonemic awareness tasks, pseudoword decoding, context-free word recognition, and reading comprehension. A two-year follow-up indicated that positive effects remained in word recognition accuracy in connected text. Finally, Hulme et al. (2012) reanalyzed data from a randomized control experiment done by Bowyer-Crane, Snowling, Duff, Fieldsend, Carroll, Miles, and Hulme (2008) using a mediation analysis and concluded that letter-sound knowledge and phonemic awareness are two causal influences on the development of children's early literacy development. Research has demonstrated that reading and understanding connected text requires the understanding of the alphabetic principle (Adams, 1990; Foorman et al, 1997; Torgesen, 2000).

Phonics

Phonics is "an approach to, or type of, reading instruction that is intended to promote the discovery of the alphabetic principle, the correspondences between phonemes and graphemes, and phonological decoding" (Scarborough & Brady, 2002, p. 20). Phonics instruction is an intentionally explicit and systematic approach to teaching decoding skills. "Explicit" refers to the direct instruction provided by teachers around key letter–sound relationships and decoding skills. Alternatively, implicit instruction would expect children to acquire these skills simply from exposure to words or incidental learning opportunities. "Systematic" refers to the structured organization of lessons and instruction. Prerequisite skills set the foundation, and a careful sequence of lessons builds to more complex skills. Phonics is not intended to teach all of the rules of phonological decoding, which number in the hundreds (Brady, 2011). However, explicit phonics instruction is meant to kick-start the process by which beginning readers acquire untaught spelling-sound relations through implicit learning (Tunmer & Arrow, 2013).

Within different phonics programs, instruction in the alphabetic code varies along several dimensions, including the degree of explicitness, systematicity, and intensity with which letter-sound relationships are taught (Brady, 2011). Phonics can differ in the size of the unit of analysis. Synthetic phonics uses the smallest units of sound for instruction, teaching grapheme-phoneme correspondences and blending skills where "children use these skills to phonologically recode unfamiliar words embodying those correspondences for themselves by translating letters into sound and blending them together" (Bowey, 2006, p. 80). Analytic phonics is a form of phonics that does not use grapheme-phoneme correspondences but focuses on patterns in words or portions of words. It introduces children to whole words before teaching them to analyze the words into their component parts, and emphasizes the larger sub-parts of words (e.g., onsets, rimes, phonograms, spelling patterns) as well as phonemes (Wyse & Styles, 2007). Teachers show children how to deduce the common letter and sound in a set of words which all begin (or, later, end) with the same letter and sound (e.g., pet, park, push, pen) (Torgesen, Brooks & Hall, 2006). Proponents of phonics claim it is useful not because of the specific letter-sound correspondences taught, which are limited in number, but because it instills in beginning readers a firm grasp of the alphabetic principle, and gives them practice in looking closely at word spelling (Snow & Juel, 2005). A key distinction in the instructional philosophy of phonics compared to other methods is the explicitness of teaching orthographic patterns and word identification strategies outside of connected text rather than teaching as the need arises during text reading. It is this structured attention to decoding using letter-sound correspondences that is the fundamental difference between it and other reading instructional methods, in particular the whole-language approach.

The whole-language approach presumes children acquire word reading knowledge and skills in large part through their experience in reading for meaning, with relatively little explicit teaching (Bowey, 2007). Constructivist in nature (Smith & Elley, 1994), the whole-language approach contends that reading will occur naturally and is the by-product of active engagement with print. It presumes "that self-discovery is the most efficacious mode of learning, that most learning can be characterized as 'natural' and that cognitive components should never be isolated/fractionated during the learning process" (Wilkinson, Freebody, & Elkins, 2000). As a result, reading evolves naturally and spontaneously out of children's pre-reading experiences with *environmental print.* Whole language theorists concluded that literacy teaching should be "modelled on first-language acquisition, where the focus is on meaning construction, not the abstract structural units that provide the basis for mapping print onto spoken language" (Ryder et al., 2008, p. 351). Instruction in grapheme-phoneme correspondences are taught implicitly, as the need arises through frequent encounters with absorbing reading materials. Minimal attention is given to the development of phonemically based, word-level skills and strategies. Central to the whole language approach is the assumption that students use multiple cues to predict the next words in text, and their meaning and word level cues are the most useful source of information when identifying unfamiliar words. While some studies have shown, depending on the beginning level of the student, both approaches can be beneficial (Connor, 2004), the majority of meta-analysis comparing the techniques have found phonics to be essential to struggling readers (Hattie, 2013; National Reading Panel, 2000; Torgesen, Brooks & Hall, 2006)

The OG Institute uses a primarily synthetic phonics approach to teach decoding skills. The full sequence of topics can be found in the appendix. The literature addressing the use of phonics in instructing children to read follows.

Evidence from the National Reading Panel. The most widely cited evidence of the benefits of phonics instruction over other instructional approaches such as the whole language approach comes from the large meta-analysis conducted by the National Reading Panel in 2000. Funded by the National Child and Human Development Department, the National Reading panel consisted of well-known reading experts who were convened to conduct a large scale review of the research on reading since 1970. The work addressed several research questions around literacy skills and instruction, including the role of phonics in early literacy instruction. Some of the research questions addressed by the panel included: "Does systematic phonics instruction help children learn to read more effectively than non-systematic phonics instruction or instruction teaching no phonics" (National Reading Panel, 2000, p. 92) and "Are some types of phonics instruction more effective than others? Are some specific phonics programs more effective than others?" (National Reading Panel, 2000, p. 93).

To evaluate these questions, the NRP conducted a meta-analysis, only including studies that had an experimental or quasi-experimental design with a control group. The studies had to provide data testing the hypothesis that systematic phonics instruction improves reading performance more than instruction providing unsystematic phonics or no phonics instruction. To be considered an instance of phonics instruction, the treatment had to teach children to identify or use symbol-sound correspondences systematically. Studies had to measure reading as an outcome and had to report statistics permitting the calculation or estimation of effect sizes. Additionally, studies were not included if they had already been used in NRP's meta-analysis of phonemic awareness training studies. Of the studies reviewed, 38 were retained to calculate effect sizes.

The results of NRP heavily favored systematic phonics over non-phonics, basal, whole-language, regular curriculum, or whole word approaches. The results of the meta-analysis were conveyed as effect sizes using *Cohen's d.* Students had better reading scores (Cohen's d = 0.44), whether they were instructed individually (.57), as a group (.43), or as an entire class (.39). The units of analysis did not differ significantly from each other compared to synthetic (.45), larger unit (.34), or miscellaneous phonics approaches (.27). The panel concluded that the biggest impact on reading growth occurs when phonics instruction begins before children are reading independently. Thus, phonics instruction should begin in kindergarten (.56) or first grade (.54) rather than in later grades (.27). The panel suggested that when taught in kindergarten, it must be appropriately designed and must begin with foundational knowledge involving letters and phonemic awareness.

> Specific systematics phonics programs are all significantly more effective than non-phonics programs; however, they do not appear to differ significantly from each

> other in their effectiveness although more evidence is needed to verify the reliability of effect sizes for each program. (National Reading Panel, 2000, p. 93)

Evidence since the National Reading Panel. Research after the National Reading panel review has further reinforced that phonics instruction is beneficial to students' reading success (Brady, 2011). In another large scale meta-analysis, Torgesen et al. (2006) was commissioned by England's Department for Education and Skills (DfES) to study the various approaches for teaching reading. Unlike the NRP, which included quasi-experimental designs, Torgesen et al. used only Randomized Control Trials in their analysis. Similarly to the National Reading Panel, Torgesen et al. concluded phonics is crucial for literacy acquisition and found there is not strong evidence any one form of systematic phonics is more effective than any other, with no difference between synthetic and analytic approaches. Torgesen et al. found only four randomized trials that tested for comprehension and no positive result was found. They determined there was insufficient evidence for whether phonics increases reading comprehension or if it should be used for spelling. Their recommendation was that systematic phonics should be a routine part of literacy teaching and used for both normally developing children and those at risk of reading failure.

More recently, Brady (2011) reviewed the research findings on the efficacy of phonics instruction in the decade following NRP. Several studies conducted since NRP supported the role of phonics as a critical element of early literacy instruction. A 2004 study by Johnston and Watson, conducted in Scotland, compared synthetic phonics and analytic phonics. The synthetic phonics group had better results, even though students in this group had more deprived backgrounds. The unit size and explicitness of phonics instruction were investigated by Christensen and Bowey (2005), using three different types of instruction. One group of students was taught analytic phonics with explicit instruction using *orthographic rimes* (OR). A second group was taught synthetic phonics using *grapheme-phoneme correspondences* (GPC). The third group was taught phonics implicitly. Roughly 40 students were in each group in the study, matched on pretests scores on phoneme segmentation, reading, and spelling scores. No differences were found in initial groups. Results were that both explicit groups performed better on accuracy in which they read and spelled words in the program. Some results favored the GPC group. The GPC group read more quickly.and only the GPC group spelled transfer words accurately. Both explicit programs had higher scores on reading comprehension and oral reading posttest than the implicit group. The authors concluded:

> In summary, this study showed that programs designed to provide explicit practice in the use of letter-sound relationships to decode unfamiliar words can significantly enhance children's performance across a wide range of measures of reading and spelling. Additionally, for relatively early readers, a program that focuses their attention on individual phoneme-grapheme correspondences and encourages them to analyze every grapheme in a word is superior to the one that encourages them to focus on larger orthographic units, specifically rimes. (p. 347)

While the research evidence has favored phonics over whole-language learning there is some evidence that different instructional approaches may benefit different types of students. Connor, Morrison, and Katch (2004) investigated the interaction between first graders' beginning decoding and vocabulary, instruction type, and end-of-the-year decoding skills. The found that children with poor decoding skills and vocabulary do better with teacher-centered explicit instruction. Students who entered with high decoding skills and vocabulary did better with a more student-centered implicit approach. Similar results had been found by Juel and Minden-Cupp (2000) who concluded "Yet, the promotion of balanced instruction leaves open the question of what might be the best combination of basic skills instruction and meaningful reading activities" (p. 306).

The research on phonics has suggested that a range of ways of teaching systematic phonics is effective (Wise & Goswami, 2010). Furthering the evidence on the benefits of phonics instruction for struggling readers, Hattie (2013) reported stronger effects for phonics instruction. The study included 14 meta-analyses of phonics that included 12,000 students in the study. The mean effect size for phonics was of 0.6 over all meta-analyses. Hattie also did a review of whole language instruction. This review included four meta-analyses on whole language involving 630 students. Almost no effect was found with a mean effect size of 0.06.

Intervention Structure. Many students fall behind, even when core instruction is of high quality. Several best practices have been proposed for how to structure interventions for students that have not been successful with core instruction. A popular current practice is to provide support through different 'layers' of intervention or 'tiers', as in the Response to Intervention (RtI) model. Although the OG institute is not part of a fully developed RtI model within the district, its development was designed with the best practices of RtI in mind. The Institute of Education Sciences (IES) released a report in

2009 titled *Assisting Students Struggling with Reading: Response to Intervention and Multi-Tier Intervention in the Primary Grades* (Gersten et al, 2009). Several recommendations for Tier II reading interventions, based on a meta-analysis of several well designed research studies, were made by the authors. Based on these works, the IES recommends that reading interventions "provide intensive, systematic instruction on up to three foundational skills in small groups of students who score below the benchmark on universal screening. Typically, these groups meet between three and five times a week for 20 to 40 minutes" (p. 19). The structure of the OG Institute aligns to these suggestions.

Text support in Phonics Programs

The last lesson in the ten lesson rotation for the OG Institute is the use of decodable books, which provides a structured means for students to practice decoding skills. The books are from the Bonnie Kline series which is part of the Project Read curriculum, another adaptation of Orton Gillingham. The books are structured to reinforce and provide practice on the recently taught skills and are sequenced to follow the letters learned by students. The benefit of decodable books, or 'little books', over basal readers that have lower levels of decodability is unclear.

Jenkins, Sanders and Vadasy (2004), studied the impact of how more or less decodable texts facilitated the developmental word reading skills of at risk first graders. Seventy nine students were included in the study, with forty in the control group and the remainder in the treatment group. Both groups received tutoring in phonics and story reading. The books used in the two groups differed in their level of decodability, 85%

versus 11% for the two groups. The decodable books were consistent with the pacing of the phonics program. However, no differences in postests were found between the two groups, showing that the differences in the books did not provide different outcomes to students.

Others have found that differences in book decodability matter. Menon and Hiebert (2005) developed a model for analyzing beginning reading materials by components of linguistic content and cognitive load. The linguistic components include attention to the use of high frequency words, rimes and decodability of texts. Cognitive load refers to the occurrence of unique words in texts, referred to as density, and text-level features such as the predictable syntactical and story patterns. The researchers compared the results of students using a traditional set of texts to another set of texts which they ordered and adapted to be leveled according to linguistic content and cognitive load. Children in the intervention group scored better on reading word lists and passages than the comparison group.

Wanzek and Vaughn (2007) found that interventions with high standardization had better results that low standardization, particularly those that emphasized both phonics and text reading. However, differences in the decodability of the texts did not seem to distinguish the effectiveness of the intervention. The research on the structure and content of texts to support phonics instruction is still mixed.

Summary of the Alignment of OG to Research on Early Literacy Skills

The instructional components of the OG institutes lesson plans align well with the research on early literacy skills, despite the fact that the program was developed prior to much of the research of the last few decades. The inclusion of lessons in phonemic awareness, the alphabetic principle and phonics align well to a large body of research showing the critical nature of these topics in developing literacy skills in young children. The inclusion of the multisensory lessons delivery is less established, although practitioners feel the activities are engaging and provide focus to children. While it is well established that young children need practice reading connected text (NRP, 2000), the exact nature of the practice and text used is less clear. However, despite the alignment to research on early literacy instruction, there is very little evidence in terms of rigorous studies in the classroom that indicates its effectiveness. As a result, a rigorous examination of its effectiveness is still needed, especially in the population of this study.

Research on the Effectiveness of Orton-Gillingham

Comprehensive Reviews of Orton-Gillingham

Several reviews of the research on the efficacy of OG and OG-based instructional programs were conducted in the last decade. The first, a literature review on OG research since 1970, was completed by Ritchey and Goeke in 2006. The authors searched the ERIC, ECER, and PsychInfo databases using the keywords: Orton-Gillingham, multisensory, reading, decoding, and phonics, as well as the titles and authors of OG-based curricula. They also reviewed dissertations conducted after 1980 as listed in ProQuest. Additionally, a hand search of popular journals on reading disability instruction including *Annals of Dyslexia*, *Journal of Learning Disabilities*, *Learning Disabilities Quarterly*, *Exceptional Children*, and *Journal of Special Education* was completed. The only studies the authors included were those published in a peer-refereed journal or were a doctoral dissertation, studied a multisensory reading approach based on OG methods or philosophy, were conducted using experimental, quasi-experimental, or single-subject research designs, and included a sample size of at least 10 participants per experimental condition. Their review found only twelve studies, including two dissertations, that met their standards for effective design. Three of the twelve studies were conducted in classrooms containing kindergarten or first grade.

The first study was conducted by Litcher and Roberge in 1979 which analyzed an initiative titled the *High Risk Experimental Project*. The study took place in seven Winston-Salem schools. Two housed the experimental classrooms and the remaining schools held the control classrooms. For three consecutive years the researchers conducted the same research project, selecting 10 matched pairs of first grade students at risk of reading failure to receive instruction in either the experimental or control classrooms. The students in the experimental classroom received Orton-Gillingham based instruction three hours a day and were not introduced to the default basal based reading program taught in the control classrooms. Students were assessed using the Metropolitan Achievement Tests on word knowledge, word analysis, reading and total reading subtests. Additionally, they were assessed on the vocabulary and comprehension subtests of the MacGinitie Reading Test. The students in the experimental group had superior scores on all subtests compared to control students, with the effect sizes ranging from 0.75 to 2.88 for different years. While the effects are quite large, the study has several shortcomings. The authors did not include any description of how the pairs were matched and presented no baseline data or demographic information on the sample.

Also, the study did not include fidelity checks on implementation and the analysis did not control for teacher effects. Furthermore, no details of the curriculum were provided other than students "used numerous teacher-made and commercial materials and based largely on the 1964 OG publication of A Guide to Teaching Phonics" (Litcher & Roberge, 1964, p. 240).

The second study examined the effects of Project Read®, a multisensory language arts curriculum based on Orton-Gillingham and designed for use in a classroom or group setting (Stoner, 1991). Two main objectives of Project Read® are to use language in all its forms, and to use responsive instruction rather than pre-planned textbook lessons. The program emphasizes direct instruction, and lessons move from letter-sounds to words, sentences, and stories (WWC, 2010). Subjects were first, second and third graders in a midwest elementary school who received instruction by district employed classroom teachers trained in the Project Read curriculum. The study utilized the "recurrent institutional cycle design" where the control group comprised students enrolled at the school in the year prior to the implementation of the Project Read® curriculum. Approximately 40 students were in each of the control and experimental groups. Equivalence between the experimental and control groups was determined by IQ scores, which were equivalent at the 0.05 level. The study only found significant gains for students in first grade. Using results on the Stanford Achievement Tests, the effect sizes were significant for Project Read® Students in Word Study Skills ($d = 1.15$), Word Reading ($d = 1.06$), Reading Comprehension ($d = 0.93$), and Total Reading ($d = 1.15$). When the authors only used teachers that taught both curricula in their analysis, the findings were the same. However, the treatment and control groups were only matched on IQ, not on reading ability, so no statistical adjustments for group differences were made.

The last study compared an OG-based curriculum called Language Basics (modeled after Alphabetic Phonics) to the Houghton-Mifflin Basal Reading Program used as the standard curriculum in the district (Joshi, Dalhgren & Boulware-Gooden, 2002). Two classrooms were taught Language Basics in one school while two classrooms in another school were used as the control. The sample included 24 students in the experimental group and 32 in the control group who received the full treatment over the course of a year. Both the experimental and control classes received 50 minutes of literacy instruction a day. The same tests were given as both pre and post measures and average gain scores on each measure were analyzed using repeated measures multivariate analysis of variance. Students were assessed on the Test of Phonological Awareness (TOPA), the Woodcock Reading Mastery Test-Revised (WRMT-R) Word Attack Subtest, and the comprehension subtest of the MacGinitie Reading Test (GMRT). Instruction was delivered by classroom teachers who received 42 hours of training on the instructional program at the beginning of the year. The treatment group had significantly higher gains on all measures than the control group. The control group only had significant gains on comprehension, whereas the treatment group had significant gains on all measures. The authors concluded that the OG based multisensory approach used in the study was more effective than the basal instruction.

When including all twelve studies in their summary, Ritchey and Goeke concluded that the methodological rigor of future OG research must be improved. Further, they argue that a “practice to research gap” exists, considering the five decades of popular use of OG without a comprehensive study reported in a peer reviewed journal. While the evidence is weak, positive student outcomes favor OG more often than not. Of the 12 studies, five reported OG instruction to be more effective than the other instruction on all outcomes. Four studies found OG to be more effective on some, but not all outcome variables.

Four years after Ritchey and Goeke released their literature review a series of more rigorous reviews of Orton-Gillingham and Orton-Gillingham based methods were conducted by the What Works Clearinghouse (WWC), a division of the Institute of Educational Statistics. The What Works Clearinghouse is an initiative of the U.S. Department of Education’s National Center for Education Evaluation and Regional Assistance (NCEE), within the Institute of Education Sciences (IES), which was established under the Education Sciences Reform Act of 2002. The WWC does not conduct new research but vets previous research to identify and report out on which educational programs, policies or practices have been found to be effective and which are not. In this way, it acts as a ‘consumer reports’ for educational interventions.

The WWC accomplishes this by assembling a team of respected and experienced researchers from the particular field of interest to conduct the review and publish the results in accessible formats to the public. Very strict guidelines are used as criteria for acceptable studies, favoring study designs that can exhibit a causal link between the intervention and the student outcome. These designs include random control trials (RCTs), quasi-experimental designs (QEDs) and regression discontinuity designs (RDs), as well as single-case study designs (SCSD). Studies that do not use one of these designs are ineligible. To be eligible, studies must also be replicable, giving enough detail to be reproduced. After being reviewed each study is given one of three ratings: Meets WWC Group Design Standards without reservations, Meets WWC Group Design Standards with reservations, Does Not Meet WWC Group Design Standards. Only RCTs and RDs can meet the highest rating if they fulfill all of the WWC standards.

Between 2010 and 2012 the WWC conducted nine reviews covering unbranded OG and eight branded OG interventions, which in total reviewed 151 eligible studies. Branded interventions are commercial programs and products that are created by an external developer who provides technical assistance (e.g., instructions/guidance on the implementation of the intervention) and sells or distributes the intervention. Branded interventions are typically trademarked and packaged and available for distribution/use beyond a single site (WWC, 2010).

The WWC reviewed 31 studies from 1989 to 2009 on unbranded Orton-Gillingham interventions. No studies met their criteria. Four of the studies used quasi-experimental designs in which the control and treatment groups were not shown to be equivalent. Two studies only had one unit assigned to the intervention, so its effectiveness could not be solely attributed to OG. Eleven studies were not primarily studying the effectiveness of the Orton-Gillingham methodology. Eleven studies did not use a comparison group. In the remaining three studies, the treatment groups did not consist of at least 50% special education students in the intervention, an additional requirement used by the WWC in their protocol for special education interventions. Since the current study is in general education classrooms, these three studies will be reviewed further to show additional flaws that weaken their results.

The first, conducted by Hishinuma (2000), measured parent attitudes towards special education services, not an OG intervention. The second, conducted by Scheffel, Shaw, and Shaw (2008) was not published in a peer reviewed journal. It only supplied descriptive baseline statistics which did not include measures of socioeconomic background or English learner status. Additionally, it did not show equivalence of the treatment or control group and no covariates were used to control for differences in their analysis. The third study, conducted by Van Hell, Bosman, and Bartelings (2003) was not taught in English. The researchers taught spelling to Dutch students using an approach called visual dictation. Although the program used kinesthetic techniques for writing the spelling words, the letters were not explicitly taught to sounds. English has a much deeper orthography than Dutch (Bowey, 2006), which makes the extension of the results to another language tenuous.

The WWC also conducted reviews of branded OG interventions including The Spalding Method ®, Alphabetic Phonics, The Barton Method, Dyslexia, The Herman Method, and Project Read. Only two studies for Project Read met the requirements of the WWC with reservations, both of which were Masters' Theses. The first by Acalin (1995) studied kindergarten through fourth grade students in Resource Specialist Program (RSP) classrooms. Students in the treatment and control were matched on gender, grade level, ethnicity and pretest scores on the Woodcock Johnson-Revised Broad Reading scores. The treatment group received Project Read instruction while the other received instruction using Reading Recovery, another Orton-Gillingham based program. Although both groups showed significant growth from pretest to post-test, there were no significant differences in growth between the two groups. The second, conducted by Bussjaeger (1993) included fourth and fifth grade students and found Project Read to be more effective than literature based instruction used in the control classrooms.

The research conducted on Orton-Gillingham based programs in the last several decades has provided only a handful of studies that are methodologically sound. Only one was conducted in Kindergarten or 1st grade and it did not find the Orton-Gillingham based curriculum, in this case Project Read, to be more effective than the other program. While in general, the results of research on OG tend to show positive results compared to the comparisons in the study, the poor designs fail to provide conclusive evidence that OG is an effective intervention.

Recent Orton-Gillingham Research.

Since the last release the WWC intervention 2012, a small number of studies have been published that address either multi-sensory instruction or Orton-Gillingham based instruction. Unfortunately, these have not furthered the understanding of OG. Giess, Rivers, Kennedy, and Lombardino (2012) studied the effectiveness of the Barton Reading and Spelling system on ninth through eleventh grade students with reading disabilities. Although all students made gains on all measures from pre-test to post-test, the study contained only nine students and no comparison group was used. Two other recent studies did not use English as the language of instruction. Nourbakhsh, Mansor, Baba & Madon (2013) used an experimental design where one of the groups participated in a multi-sensory Orton-Gillingham approach. However, the study took place in Iran, targeting third grade students diagnosed with dyslexia. Similarly Hwee & Houghton (2011) studied the effectiveness of an Orton-Gillingham approach with Singaporean children.

Several reports, concluding in 2011, were released that studied a large multi-year research project on OG based instruction (Bitter, 2011). In this study, 11 Arizona schools participated with five serving as experimental conditions and the remaining as controls. The sample included 1000 students in 23 experimental classes and 24 control classes. A series of early literacy assessments, the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) was used as the outcome measure. Students in the treatment group had higher scores than the control group on all measures, with an average effect size of 0.54. While the study has a large size and several years of data, the treatment and control groups were not equivalent and the analysis did not control for group differences.

Summary

In sum, the preceding review highlights a lack of quality research on the effectiveness of the Orton-Gillingham method. While the method is in need of a rigorous evaluation, the research shows its instructional components have merit and align to the research on early literacy skills. Other explicit, systematic phonics programs that include instruction in phonemic awareness and the alphabetic principle have been shown to be effective. Although the multisensory components of the program do not have scientific backing to show their inclusion is necessary, or even beneficial, there is also not evidence that show they are detrimental. In fact, the advocacy for OG based methods from practitioners over the years must lend some support to their usefulness in directing a student's attention and reinforcing their learning. Most importantly, however, it is clear that OG based programs have not been well researched, especially in the target population of this paper, and evidence of the efficacy of the program is needed.

CHAPTER 3: METHOD

Introduction

This chapter will provide detailed information about the participants, procedures, measures and analysis used in the present study. Two statistical procedures on observational data will be conducted to investigate the effectiveness of the OG Institute intervention. The first analysis will employ a Regression Discontinuity (RD) design, which is appropriate because students were placed in the intervention if they scored below a predetermined cut-point on the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) assessment. The data requirements, statistical methods, assumptions and threats to validity for RD designs will be outlined. The second procedure is a Nonequivalent Comparison Group Design (NECGD) where the outcomes of similar students in control and treatment schools will be compared. Since the students were not randomly assigned to the OG Institute treatment, propensity scores will be used to match treatment and control students.

Participants

Data for this study was collected from a large urban K-12 school district in California. In the 2014-15 school year, the district had a population of approximately 43,000 students, with nearly 75% identified as ethnic minorities. The largest ethnic group was Hispanic (60.3%) followed by white (23.9%), African American (7.0%) and Asian American (4.7%). Students who identified as American Indian, Alaskan Native, Hawaiian Native, Pacific Islander or multiple races made up 4% of the population. Approximately 63% of the students in the district received free or reduced lunch, 16%

were classified as English Learners, and 10% were enrolled in Special Education programs. The population in kindergarten and first grade varied only slightly from the district totals. The largest difference was in the proportion of English Learners, where in these grades they made up 30% of the population. Of the 29 elementary schools in the district, 19 implemented the OG Institute in phases over four years. The complete demographics for each school from the district that is used in this study are outlined in Table 1.

Implementation of OG Institute

The decision to use the Orton-Gillingham method to provide reading intervention in kindergarten and first grade was made by district instructional leaders and administration. The OG Institute lessons were designed to be delivered by the classroom teacher through small group supplemental instruction to students identified as at-risk for reading failure. Although the district does not formally implement a Response to Intervention (RtI) framework, the intervention was designed with many of the components of Tier II intervention. Students were identified as being at-risk by using a universal screening measure, in this case DIBELS. Instruction was delivered to small groups consisting of no more than six students per classroom. The groups met five days a week for approximately 20 minutes per day. The aforementioned components are considered best practices for Tier II interventions based a review of the research on RtI by the Institute of Educational Science (IES) (Gersten et al, 2009).

In the spring prior to the first year of implementation (2011-12), principals were provided information on the OG institute and were offered the opportunity to participate in the first cohort of schools. Schools were selected to implement the OG Institute by district level personnel from this interested pool of schools. The schools were selected based on a combination of staff willingness to participate, readiness, and need. Because different trainers were used for different cohorts, variations in outcomes will be analyzed to investigate if this may have influenced the success of the intervention. Since the implementation of the OG Institute required in-depth training and the collective bargaining agreement in the district does not allow mandatory professional development to the extent necessary, the decision to implement had to include teacher agreement. Since schools were not selected at random, the willingness of the school may influence the intervention, which may limit the generalizability of the results. Five schools were selected for cohort 1 in the 2011-12 school year, seven schools were selected for cohort 2 in 2012-13 and seven schools were selected for cohort 3 in 2014-15. Due to a variety of circumstances in the district, including insufficient resources to provide training and support, no new schools were trained in the OG institute in 2013-14. At the time of this study, eleven schools had not participated in any of the years. The district plans on training the remaining schools in upcoming years. The number of schools in each cohort can be seen in Table 2

Classroom Instruction

Teacher Training

All kindergarten and first grade teachers at participating schools attended a five day, forty hour training over the course of one week during the summer. The principal of each participating elementary school also attended the training so they could learn the structure of the program and monitor its implementation. The training was conducted by the Institute of Multi-Sensory Education (IMSE), a private company that provides professional development in Orton-Gillingham based instruction. District reading specialists participated in the initial training and in following years provided the week long training without IMSE. Teachers who were unable to attend the summer training were required to attend make-up sessions at the beginning of the school year. Throughout the school year optional training opportunities were provided for teachers who asked for additional support. Teachers made the decision whether or not to attend the optional trainings

Lesson Plans

The lessons for the OG Institute were developed as joint effort between IMSE and district staff. The program consists of a cycle of ten lessons, which are referred to as *steps*. Two steps are to be completed each day and after the completion of the last step, the teacher begins again with step one. The titles of the ten step lesson are provided in Table 3 and a full description of the lessons can be found in Appendix A.

The scope and sequence of topics was provided to teachers for the entire year of the intervention and is available in Appendix B. The lesson plans were provided during the initial training and were also made available to the teachers online. Instructional videos were also included demonstrating the proper way to complete each lesson. District reading specialists conducted unannounced visits to each site several times during the year to check fidelity of the implementation. The fidelity checklists from the visits will be reviewed to see if substantial variation in implementation existed. If so the implementation rating can be included in the analysis as a covariate. The fidelity checklist is included in Appendix C. Site principals were also instructed to regularly visit classrooms implementing the intervention to monitor fidelity.

Selection process for students

Students were selected into the intervention based on composite score from the beginning of year (BOY) administration of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) assessments, described in detail in the next section. Students who scored *well below benchmark* were placed in the intervention. The designation *well below benchmark* is used by the authors of DIBELS to identify students who need intensive support and whose success is unlikely if only being supported by the core curriculum (Good & Kaminski, 2011). Students in the intervention were instructed by their classroom teachers within the classroom.

Although, the OG institute was designed with components of Tier II interventions, not all components were in place. Teachers were not trained on methods of identifying poor responders, such as Low Growth (Fuchs et al, 2004), Median Split (Vellutino et al (1996) or Dual-Discrepancy (Fuchs & Fuchs, 1998). Additionally, there is no Tier III intervention in place. As a result, students placed in the OG institute received the treatment for the entire school year. While this does not follow the best practices of RtI (Gersten et al, 2009), it provides more optimal conditions for evaluating the effectiveness of the intervention.

Measures

Data Collection Times

Scores from the DIBELS assessments were collected for all students in the district three times a year in September, January & May. The early literacy skills that are assessed with the different subtests in DIBELS are outlined in Table 4. Different assessments from DIBELS are given at different times of the year based on the early literacy skills children should acquire at that time. The timelines of the assessment components are provided in Table 5. Prior to the 2012-13 school year, all students in grades two and above took the California Standards Tests (CST) as required by the state and federal government as part of the No Child Left Behind Act. CST scores for the first cohort of schools are available for 2nd and 3rd grade and will be used to investigate the long term effect the intervention.

Dynamic Indicators of Basic Early Literacy Skills (DIBELS)

The Dynamic Indicators of Basic Early Literacy Skills (DIBELS) is a set of measures used to assess early literacy and reading skills for students from kindergarten through sixth grade (Good, Kaminski, Cummings, Dufour-Martel, Peterson, Powell-Smith & Wallin, 2011). DIBELS can be used to identify students who may be at risk for reading difficulties and provides on- and off-track indicators of students' reading ability (Good et al, 2011). Kindergarten students are assessed on letter knowledge, phonemic awareness and the alphabetic principle. First grade students are assessed on phonemic awareness, the alphabetic principle, phonics and reading fluency. All tests are administered individually.

On each assessment of DIBELS students are identified as having one of three levels: *At or Above Benchmark*, *Below Benchmark*, and *Well Below Benchmark*. Benchmark goals for DIBELS are based on research that examines the predictive validity of a score on a specific measure at a particular point in time, compared to later DIBELS measures and external outcome assessments (Good et al, 2011). If a student achieves *At or Above Benchmark* goal, then the odds are in favor (80-90%) of that student achieving later reading outcomes if the student receives research-based instruction in core classroom instruction. Students that score *Below Benchmark* are unlikely to achieve subsequent reading goals without receiving additional, targeted instructional support and students that score *Well Below Benchmark* are likely to need intensive support. Students who require intensive support meet later benchmark only10-20% of the time, unless supplemental instruction is provided (Good et al, 2011). The assessments are created using grade-level material. The measures administered for benchmark assessment vary by grade and time of year, and include those measures that are most relevant for making instructional decisions at that time. The cut-points for risk categories are nationally normed, empirically derived, criterion-referenced target scores that represent adequate reading progress (Good et al, 2011). All DIBELS assessments were individually administered by the student's classroom teacher. Prior to the first year of implementation of the OG Institute, all teachers were trained in the administration of DIBELS. Descriptions of the individual assessments follow.

First Sound Fluency (FSF)

First Sound Fluency (FSF) measures a student's ability to identify the initial sounds in words. The assessor says a series of words one at a time to the student and asks the student to say the first sound in the word. Three practice items are included to ensure that students understand the task and to maximize the performance. Each form consists of 30 items, including words that have initial continuous sounds, initial stop sounds, and initial blends. A student's score is the correct number of first sounds spoken from this list in one minute.

Letter Name Fluency (LNF)

Letter Naming Fluency (LNF) measures a student's ability to recognize letters and say their names. Students are presented with a page of uppercase and lowercase letters arranged in random order and are asked to name the letters. On an identical paper held out of view of the student, the assessor marks letter names that are read incorrectly or skipped. The total score is the number of correct letter names that the student says in one minute. The purpose of LNF is to measure fluency rather than identify which letters the student knows or does not know, so while all letters are included on the LNF materials, they appear in random order. As such, it provides an added risk indicator for early school-age children. Because letter naming by itself has been shown to not be essential for achieving reading outcomes (Blachman et al, 1996), a benchmark goal is not provided. Scores on LNF are used as part of the student's composite score.

Phoneme Segmentation Fluency (PSF)

Phoneme Segmentation Fluency (PSF) measures the student's ability to segment words into its component phonemes. The assessor says a word and asks the student to say all the sounds in the word. PSF includes 3- and 4-phoneme words. For example, if the examiner says "dog" the child needs to say d/o/g. A student receives one point for each correctly spoken phoneme, in the case of the example they would receive three points. The assessment has 24 items and the final score is the number of phonemes correctly identified in one minute.

Nonsense Word Fluency

Nonsense Word Fluency (NWF) measures the student's ability to read nonsense words. The student is presented with a list of VC and CVC nonsense words (e.g., tuf, id, wit) and asked to read the words. The assessment is used to measure the student's use of alphabetic decoding. There are two separate scores reported for NWF. Correct Letter Sounds (CLS) is the number of letter sounds produced correctly in 1 minute. Whole Words Read (WWR) is the number of make-believe words read correctly as a whole word without first being sounded out.

For example, if the student reads wid as "wid," the score is 3 points for CLS and 1 point for WWR, but if the student reads wid as "/w/ /i/ /d/ wid," the score is 3 points for CLS but 0 points for WWR. The goal is for students to read whole words on NWF. An advantage of NWF, however, is that it also provides a measure of the student's development of the alphabetic principle and basic phonics.

DIBELS Oral Reading Fluency

DIBELS Next Oral Reading Fluency (DORF) measures the student's ability to read accurately and fluently using grade-level text. Each DORF passage is a 1-minute long fluency measure. Students are asked to read 3 passages aloud for 1 min each. Two scores are collected by the teacher. The first is the number of words correctly read within the one minute assessment. The second is the number of errors that the student makes during the one minute assessment. From those two measures, the student's accuracy is calculated by taking the number of words correct divided by the number of words attempted (i.e. the number of words correct plus the number of errors).

DIBELS Composite Score

Based on a combination of the scores from each administration, a composite score is created to provide an overall indicator of the student's ability level. The DIBELS Composite Score is a combination of multiple DIBELS scores and provides the best overall estimate of a student's reading proficiency. The scores used to calculate the composite vary by grade and time of year, so composite scores should be compared to the benchmark goal only at that time of the school year, and not to composite scores at other times of the year. Scores at or above the goal indicate that the student is on track for meeting future reading outcomes. For example, in kindergarten, the composite score at the beginning of the year is the total of FSF and LNF scores and at the middle of the year it is the total of FSF, LNF and PSF scores.

Analytical Method

Analysis 1: Regression Discontinuity Design

A regression discontinuity design, if implemented correctly, can provide unbiased estimates of the treatment effect of an intervention similar to a randomized experiment (Bloom, 2009; Lee & Lemieux, 2010; Jacob, Zhu, Somers & Bloom, 2012). Regression Discontinuity was first introduced by Thistlethwaite and Campbell (1960) in their study of the effects of merit awards on future academic outcomes. While its popularity diminished over the next several decades, use of Regression Discontinuity has recently been revived (Shadish, Campbell & Cook, 2002). Recently, RD designs have been used in educational analysis to study financial aid offers on college enrollment (van der Klaauw, 2002), remedial education and grade retention (Jacob & Lefgren, 2004), mathematics interventions for first- and second- grade students (Bryant, Bryant, Gersten, Scammacca & Chavez, 2008), and a Tier II reading intervention in first grade (Baker, Smokowski, Chaparro, Smith & Fien, 2015).

Situations that lend themselves to an RD approach occur frequently in practice and one can often obtain existing data and use it post hoc to conduct analysis of program impact (Jacob et al, 2012), which is how RD will be used in the present study. In an RD design, an ordinal variable is used to assign units to a treatment condition. The observed variable, $\boldsymbol{X}$, is referred to as the forcing or rating variable. A score below or above a predetermined cut-point, ***c***, on this variable is used to assign subjects to either a treatment or control condition. The basic assumption of the RD design is that, in absence of treatment, a smooth function exists between the rating variable ***X*** and the outcome variable ***Y***. If a treatment is effective a jump, or discontinuity, at the cutpoint will occur providing a direct measurement of the treatment effect. Figure 3 gives an example of a relationship between the rating variable and the outcome with and without treatment.

Theoretically, RD designs should produce unbiased estimates similar to a Randomized Controlled Trials (RCT) (Trochim, 1984). Several empirical studies have tested this claim by simultaneously conducting an RCT and RD design on the same program, finding equivalent results for both analyses (Aiken, West, Schwalm, Carroll & Hsiung, 1998; Buddelmeyer & Skoufias, 2004; Black, Galdo & Smith, 2007; Berk, Barnes, Ahlman, & Kurtz, 2010; Cook, Shadish & Wong, 2008). This study will follow the suggested steps for an RD analysis as outlined by Jacob, Zhu, Somers & Bloom (2012) in *A Practical Guide for Regression Discontinuity*. The guide was created by leading researchers who have used RD designs in their published works. It does not provide new research on RD, but offers a series of best practices based on the developments of recent research. The guide includes tests for checking threats to internal validity, how to correctly specify the functional form of the relationship between rating variable and outcome, and estimation of the treatment effect.

Internal Validity

The internal validity of a research design enables inferences to be made about the causal relationship between the treatment and the outcome (Shadish, Cook & Campbell, 2002). Fundamental to the internal validity of an RD study is that a candidate's ratings and the cut-point must be determined independently of each other (Bloom, 2009). Jacob et al. (2012) recommend four techniques that provide evidence that manipulation of the ratings process did not occur. First, the process used to establish the rating variable and determine the cut-point for assigning subjects to treatment should be well documented, outlining how the cut-point was determined and how subjects were assigned to treatment. Second, baseline covariates should be examined to see if any discontinuities exist at the cut-point. Third, the density of the rating variable should be examined at the cut-point. Fourth, a McCrary test for discontinuity should be used on the rating variable at the cut-point.

The Ratings Process. Students were placed into the OG Institute if they scored below a cut-point on the composite score of the DIBELS assessments administered in September. In kindergarten, two assessments are given: First Sound Fluency (FSF) and Letter Naming Fluency (LNF) (see Table 5). The composite score is created by adding the two measures together. Students were placed in the OG Institute if they scored less than 13 on their composite score. The use of this cut-point follows the guidelines provided by the authors of DIBELS (Good et al, 2011). Based on nationally representative samples from the United States, only 10-20% of students in this score range make future reading targets (Good et al, 2011). Similarly, in first grade the students were given three assessments: Letter Name Fluency (LNF), Phoneme Segmentation Fluency (PSF), and Nonsense Word Fluency (NWF). Students with a composite score of 97 or lower were assigned to the intervention. There is little or no incentive for the teachers to adjust student scores, so manipulation is unlikely. Although DIBELS has been administered in the district for several years, and all teachers have been trained on its administration, interrater reliability data is unavailable. It is also important that the forcing variable not be used to assign subjects to other interventions (Jacob et al, 2012). In the case of this study, students in the OG institute received no other reading interventions other than the OG Institute.

To provide further evidence of the ratings process, a graph showing the probability of receiving treatment should be examined (Jacob et al, 2012). In a 'sharp' RD design, every subject on the side of the cut-point that is eligible for treatment would receive it, and all those on the other side would not. This is equivalent to the probability of receiving treatment to be one on the treatment side of the cut-point and zero on the non-treatment side. In certain cases, such as when resources are limited, not all subjects that qualify for treatment receive it. Similar to RCTs, these subjects are considered 'no shows' (Bloom, 2009). In the present study, the intervention size, which was limited to six students per class, placed a constraint on being able to provide treatment to all students. In some classrooms, where more than six students were eligible, not all received treatment. In contrast to a sharp RD design, 'fuzzy' RD designs denote noncompliance in subjects to either treatment or control. A fuzzy RD design is still valid, but requires a statistical adjustment to recover the treatment effect. In the present study, it is approximated that 12% of eligible students did not receive the intervention due to the limitation in the group size of the intervention. There were no cases when less than six students qualified for the intervention. The adjustment required to recover the treatment effect when not all eligible subjects receive it will be discussed later.

Examining Non Outcome Variables. Lack of manipulation of the ratings process can be further supported through the analysis of baseline covariates and prior achievement. Baseline covariates or prior student achievement should not be affected by the treatment (Jacob et al, 2012). A discontinuity at the cut-point on these variables could signal a manipulation of treatment and render the analysis invalid. A graph will be created of each baseline covariate against the rating variable. First grade students will also have prior year DIBELS results analyzed, as they should not be affected by treatment a year later. In each case, a graph without any 'jumps' in the distribution at the cut-point would strengthen the argument that manipulation did not occur.

Density of the Ratings Variable. Additionally, it is suggested by Jacob et al (2012) to plot a histogram of the rating variable at each value. If the RD design is valid, there should be no discontinuity observed in the number of observations just below and just above the cut-point. A jump at the cut-point may indicate manipulation of the ratings variable. For example, if student scores on an exam granting access into a highly prestigious program, there could be motivation to give students who scored just below the cut-point a ‘boost’ in their score to make them eligible. A graph of the composite scores of DIBELS at the beginning of the year will be analyzed for a discontinuity at the cut-point. A discontinuity at the cut-point would question the validity of the results.

Aside from visual inspection of a discontinuity, an empirical test is also suggested. McCrary (2008) created a test to assess the discontinuity of the ratings variable at the cut-point. The McCrary Test is performed by creating a histogram of the ratings variable using equal size *bins* that do not overlap the cut-point. The bins are created by grouping a collection of points that are near each other into a single point. Regressions are calculated for both sides of the cut-point using these bins. In the regression equations, the midpoint of the bin is used as the independent variable and the bin frequency count is the dependent variable. The log difference in heights between the intercepts from the two regression lines is tested for significance from zero.

Model Specification

Proving that the ratings process was not manipulated is critical, but by itself does not guarantee an unbiased estimate of the treatment effect. To provide an unbiased estimator of the treatment effect, choosing the correct functional form between the rating variable and the outcome variable is required (Jacob et al, 2012). Choosing this correct functional form is considered the most challenging aspect of RD designs (Lee & Lemieux, 2010). Figure 4 illustrates how a misspecified model can falsely attribute a treatment effect to an intervention. In this case, the correct form is nonlinear but a linear model was fit to the data, showing a discontinuity at the cut-point when one does not exists. In most cases, the functional form is unknown and requires several empirical tests to validate it (Jacob et al, 2012). Two approaches, parametric and nonparametric, can be used to specify the functional form between the rating variable and the outcome (Bloom, 2009).

The Parametric Approach. The parametric approach uses the entire range of observations in the model. In this regard, it "borrows strength" (Jacob et al, 2012) from observations far from the cut-point. Many different functional forms should be tested, from linear to higher order forms including interaction terms. To establish which model is the best, F-Tests are conducted on the R-squared statistics of the models to establish improved fit.

Nonlinear relationships between outcomes and ratings are often explained by two theoretical reasons (Jacob et al, 2012). The first is that a ceiling or floor effect on the outcome may exist. The second is that the treatment effect may vary systematically with treatment. Since the relationship between the outcome and the rating variable are rarely supported by theory, many functional forms should be tested to find the model that best fits the data (Lee & Lemieux, 2010). Equation 1 shows the simplest linear form that an RD design can take.

$$Y_i = \alpha + \beta_0 \cdot T_i + f(r_i) + \varepsilon_i$$

$Y_i =$ Outcome for subject i;
$\alpha =$ the average value of the outcome for those in the treatment group after controlling for the rating variable;
$\beta_0 =$ the marginal impact of the program at the cut-point;
$T_i =$ 1 if observation i is assigned to the treatment group and 0 otherwise
$f(r_i) =$ the functional form of the rating variable and the outcome. Several forms should be investigated both with and without interaction terms.
$r_i =$ the rating variable, usually adjusted to be 0 at the cut-point.
$\varepsilon_i =$ a random error term for observation i, which is assumed to be independent and identically distributed.

Equation 1: A linear parametric model. From Jacob et al (2012)

Lee & Lemieux (2009) suggest testing six functional forms: linear, linear interaction, quadratic, quadratic interaction, cubic, and cubic interaction. In a recent study that investigated the effect of a Tier II intervention on reading achievement (Baker et al, 2015), the researchers used data from control schools to model functional form of the relationship between the ratings variable and the outcome variable. As noted by the authors, this approach is often not available in many RD designs, since the population of the study typically consists of only the population that were used to place subjects into the intervention or not. However, this approach is also available in the present study, since many schools administered the same assessments to all students, but did not implement the intervention. This allows for an additional level of certainty about the correct functional form, since it can be studied in a population in the absence of treatment.

The Nonparametric Approach. The nonparametric approach narrows the range of observations to be within close vicinity of the cut-point and local linear regression is applied the regions above and below the cut-point. In this approach, the model is specified to be linear and the slope can vary on either side of the cut-point. In the parametric approach, all the data is used to specify the correct functional form. Conversely, in the nonparametric approach, the data is truncated to fit the specified linear model around the cut-point. Since the data is restricted to an area close to the cut-point, bias can result. It has been advised to use a local linear regression to reduce this bias, where the slope and intercept are allowed to differ on either side of the cut-point (Hahn, Todd, and van der Klaauw, 2001).

The empirical challenge in the nonparametric approach is selecting the correct bandwidth for the analysis, as an incorrect bandwidth can bias estimation (Lee & Lemieux, 2008). Two methods, the Cross-Validation Procedure (Ludwig & Miller, 2005; Imbens & Lemieux, 2008) and the "Plug-In" Procedure (Imbens & Kalyanaraman, 2009, DesJardins & McCall, 2008), are suggested by Jacobs et al. (2012).

The Cross-Validation procedure uses a "leave-one-out" process, where a regression is estimated within a bandwidth with a single point at the end of the bandwidth removed from the analysis (Imbens & Lemieux, 2008). This point is then predicted by the regression line to determine the accuracy of the model within the bandwidth. This is done iteratively on both sides of the cut-point, shifting the bandwidth one point at a time. The error terms are used to determine the accuracy of the bandwidth. This is done for varying bandwidth sizes, with the bandwidth that produces the lowest error term being considered the best choice.

The Plug-in procedure uses a formula developed by Imbens and Kalyanaraman (2009) and DesJardins and McCall (2008) that was created to balance bias and precision. It is also an iterative process that is computationally demanding, but software has been created by Imbens that is available online. Both analyses will be conducted to investigate a nonparametric model although it has been shown that both provide similar results (Gleason, Resch, and Berk, 2012). Once a bandwidth has been determined, local polynomial regressions will be tested to assess if the linear model is overly restrictive. Jacob et al. (2012) suggest that results from both the parametric and nonparametric approach can be used as a sensitivity analysis. If results are stable across all plausible specifications, then they can be considered more robust and reliable than those that are extremely sensitive to specifications.

Estimating the Treatment Effect

In a sharp regression discontinuity design, the jump at the cut-point provides a direct measure of the treatment effect. As shown in Equation 1, the magnitude of β_0 provides this measure. In a fuzzy RD design, this estimate needs to be adjusted since the probability in receiving treatment does not change by a value of one at the cut-point. Bloom (2009) shows that since it is reasonable that no-shows receive zero treatment effect as a result of assignment, the treatment effect in an RD design with no-shows is the weighted mean of the average effect at the cut-point of those receiving treatment and the zero effect for no-shows. This can easily be computed by taking the treatment effect and dividing it by the change in probability of receiving treatment. In the case of this study, the treatment effect, which would be diluted slightly due to no-shows, would be divided by approximately 0.88, since about 12% of the eligible population did not receive treatment. No students received the OG intervention if they did receive a score below the cut-point.

Heterogeneity of Treatment Effects

The regression discontinuity analysis will show if a treatment effect exists for participants pooled across all schools and classrooms. Even if an overall effect exists, there may still exist variability in the effectiveness in the intervention between clusters. If there is significant variance between classrooms or schools it signifies possibility that the effect might not be significant in some units. Furthermore, such variance would indicated that characteristics of the school or classroom may be related to the effectiveness of the intervention, which would have consequences for successful replicability of the intervention.

Similar to Baker et al (2015), the heterogeneity of treatment effects will be tested by using a series of Hierarchical Linear Models (HLMs). Two models will be used, one with the level of clustering at the classroom level, and the second with the level of clustering at the school level. The equations in Equation Set 1 will be used in both analyses, with level 2 representing either the school or classroom levels. β_{oj} represents the intercept, or average DIBELS end-of-year achievement at level 2. β_{1j} represents the treatment effect, or the average added benefit a child would receive by being placed in the OG Institute. A significant variance in u_{ij} , which is the school or classrooms unique contribution to the OG treatment effect, signifies heterogeneity of treatment effects across the level 2 cluster (either the classroom or school). The variables $\beta_{qj} = \gamma_{0q}$ for q = 2,3,…,13, are student level controls that adjust for differences in student background characteristics. If the variance in u_{ij} is significant, classroom or school covariates will be added to the level 2 equation to investigate if the variation in treatment can be reduced by school or classroom level variables. For example, the proportion students with EL or poverty status in the classroom or school might account for variation in treatment effects across units. The objective of this examination is to gain insight into what is causing that variation and to inform practice in terms of the conditions under which one can expect the treatment to work.

Equations Set 1:

Level 1 Equation

$$Y_{ij} = \beta_{oj} + \beta_{1j}(OG)_{ij} + \beta_{2j}(DIBELS_{BOY})_{ij} + \beta_{3j}(NSLP)_{ij} + \beta_{4j}(HSG)_{ij} + \beta_{5j}(SOMECOLL)_{ij} + \beta_{6j}(COLLGRAD)_{ij} + \beta_{7j}(FEMALE)_{ij} + \beta_{8j}(EL)_{ij} + \beta_{9j}(SWD)_{ij} + \beta_{10j}(WHITE)_{ij} + \beta_{11j}(ASIAN)_{ij} + \beta_{12j}(AA)_{ij} + \beta_{13j}(OTHERETHNICITY)_{ij} + r_{ij}$$

Level 2 Equation

$$\beta_{oj} = \gamma_{00} + u_{oj}$$
$$\beta_{1j} = \gamma_{01} + u_{1j}$$
$$\beta_{qj} = \gamma_{0q} \text{ for } q = 2,3,\ldots,13$$

Analysis 2: Propensity Score Matching

The second analysis will use a Nonequivalent Comparison Group Design (NECGD) where propensity scores will be used to equalize group differences. In Randomized Controlled Trials, treatment and control groups will not differ on observed and unobserved baseline covariates (Rubin, 1974). However, in observational studies when treatment is not randomly assigned, treatment and control groups can often differ systematically on baseline covariates (d'Agostino, 1998). In such cases, propensity scores can be used to balance the treatment and control groups on all measured covariates (Shadish & Steiner, 2011). When proper conditions are met, the use of propensity scores in quasi-experimental studies have been shown to produce unbiased estimates of the treatment effect (Dehejia & Wahba, 2002; Shadish, Clark, Steiner & Hill, 2008) and closely approximate those of RCTs (Becker & Ichino, 2002). Propensity scores were defined by Rosenbaum and Rubin (1983) as the expected probability of treatment assignment conditional on observed baseline covariates: $e_i = Pr(Z_i = 1 \mid \boldsymbol{X_i})$ where Z_i = 1 is the probability that individual *i* is in the treatment group (Z = 0 for control and Z = 1 for treatment) and *X* is a vector of all covariates. Since propensity scores represent a probability, they fall between 0 and 1. They are most commonly estimated using a logistical regression, where all observed covariates are regressed on the outcome of being in control (Z = 0) or treatment (Z = 1) (Shadish & Steiner, 2011). In a group of subjects who all have the same propensity score, the distribution of observed baseline covariates will the same for treatment and control groups (Austin, 2011).

Propensity scores exist in both randomized and observational studies. In randomized studies, all individuals are equally likely to be placed in either group so all individuals have a propensity score of 0.5. In observational studies, the propensity scores are unknown and must be estimated (Shadish & Steiner, 2011). Once balanced groups are created using propensity scores, the effect of the treatment can be computed by taking the difference in mean outcomes between treatment and control groups, just as would be done in a randomized control trial (d'Agostino, 1998). An overview of the procedure follows.

Creating Propensity Scores

Propensity scores can be created using several different methods such as logistical regression, bagging or boosting (Lee, Lessler, & Stuart, 2010), tree-based methods (Lee et al, 2010) and neural networks (Setoguchi, Schneeweiss, Brookhart, Glynn & Cook, 2008). Most often in practice, logistical regression is used to estimate propensity scores

(Austin, 2011) and has been found to produce unbiased estimates of the treatment effect if the correct covariates are chosen (Shadish et al, 2008, Thoemmes & Kim, 2011).

To create propensity scores using logistical regression, the baseline covariates, represented as a vector X, are regressed onto the outcome Z, where Z = 1 for the treatment condition or Z = 0 for the control condition. Propensity scores can be created using stepwise regression beginning with all covariates and removing them, or beginning with a single covariate and adding them (Austin, 2011). Shadish and Steiner (2010) propose beginning with a fully specified model and including all covariates, even if certain ones are not significant in the model. The covariates that will be used in the model can be seen in Table 6.

The choice of covariates plays a crucial role in the calculation of propensity scores (Steiner & Cook, 2013). The critical to the internal validity of a research design using propensity scores is the concept of *strong ignorability* (Rubin & Rosenbaum, 1983). Strong ignorability is defined as the absence of an unobserved covariate that is confounded with treatment given the observed covariates (Rubin & Rosenbaum, 1983). Treatment assignment is considered to be strongly ignorable if the treatment assignment, Z, and the response, Y, are known to be conditionally independent given the covariates, X. Strong ignorability is violated when a variable that is correlated with treatment is not measured and included in the model. Violating strong ignorability introduces bias into the estimate of the treatment effect. There is no empirical test for strong ignorability (Cook & Steiner, 2010). However, the inclusion of several covariates that are correlated with the outcome can reduce the threats to strong ignorability and reduce the bias of the treatment estimate (Cook & Steiner, 2011).

The inclusion of a pretest measure to calculate propensity scores can greatly reduce the bias of the estimate (Shadish et al, 2008). The use of DIBELS is not a pretest per se, as each administration of DIBELS assesses developmentally appropriate skills. However, studies have shown that composite scores on an earlier administration of DIBELS are highly correlated and predictive of later scores (Good et al, 2001). Additional to the pretest measure, several other covariates associated with achievement will be included. Family background has been shown to be highly correlated with student achievement (Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld & York, 1966; Duncan, Brooks-Gunn, & Klebanov, 1994). Two measures of family background will be used as covariates: Parent education and participation in the National School Lunch Program. Additionally, measures of English language status, ethnicity and gender will be used. While there could be other selection mechanisms at work, the use of several correlated covariates with outcomes have shown to reduce all if not most of the bias in treatment effects (Shadish et al, 2008).

Creating Groups and Assessing Balance

Treatment and control groups can be created using propensity scores in many ways. One-to-one matching, in which a single control subject is matched to a treated subject, is popular but perhaps not most efficient (Steiner & Cook, 2013). In one-to-one matching a single treatment subject is matched to a control subject with an equal (or near equal) propensity scores. Several variations exist, including matching with or without replacement, greedy matching, nearest neighbor and other approaches (Austin, 2011).

While one-to-one matching has been a popular method for creating groups with propensity scores, Cochran (1968) found that the use of five stratum can reduce the bias treatment effects. In the stratification approach, individuals from control group and treatment group or ranked by propensity score in their respective groups. The two groups are then divided into strata based on a range of scores, with five strata being suggested as sufficient (Cochran, 1968). Although 7 – 10 strata have been used (Hong & Raudenbush, 2008), increasing strata size quickly diminishes returns and can create non-overlapping groups (Cook & Steiner, 2010). In an analysis of five stratum that includes all possible propensity scores, the first stratum would consist of subjects with propensity scores that range from 0.80 to 1.00. The second stratum would consist of subjects between 0.60 and 0.79 and the remaining stratum would be created in similar fashion. This study will use a minimum of five strata in that analysis of the treatment effect. The control group will be created from students at schools within the district with similar demographics that did not implement the OG Institute.

After Propensity scores have been calculated and groups created, the resulting balance between control and treatment groups should be reviewed in several ways (Thoemmes & Kim, 2011). Balance refers to the control and treatment groups having almost identical distributions on all pretreatment covariates, with the goal to mimic pretreatment covariate balance in a randomized experiment (Cook & Steiner, 2010). To assess balance, a visual inspection of the distributions of the covariates should first be examined for each covariate between the treatment and control group. This can be done through the comparison of histograms, kernel density estimates or QQ-plots (Steiner & Cook, 2013). After the visual inspection, the balance can also be examined empirically in a few different ways. Rubin (2001) suggests three additional empirical tests. The first is examining the standardized mean difference (*d*) in the propensity scores between control and treatment groups, where *d* should be close to zero ($|d| < 0.1$). The second is testing that the ratio of the variances on the propensity scores between the control and treatment groups is close to 1. The third is assessing that the ratios in the covariates after adjusting for the propensity scores is close to 1.

Once propensity scores are calculated, the treatment effect can be calculated by looking at the mean difference in outcomes between treatment and control groups. Only the region of common support, where there is overlap between the control and treatment groups with equivalent propensity scores, should be used in the analysis (Thoemmes & Kim, 2011). This will limit the generalizability of the results to students likely to be placed in the OG Institute, but the intervention was designed to assist these students, so this restriction does not affect the usefulness of the findings. When strata are used, as in this study, the average difference between each strata is calculated to create the treatment effect for the intervention.

Expected Outcomes of the OG Institute

The literature review presented in this manuscript provides evidence that the instructional components of the OG Institute align well to the research on early literacy. Given this evidence, it is reasonable to expect that the OG Institute would provide a positive effect for students placed in the intervention. While previous OG based interventions, given their methodological shortcomings, have provided little in terms of expected outcomes, many reviews of similar early literacy interventions provide insight into the range of possible effects a robust intervention may produce, to which the OG Institute should be compared.

It is important to put the expected effect of the intervention into perspective for a number of reasons. First, several early reading interventions, as indicated in Chapter 1, have increased the reading achievement of struggling readers. If the OG institute produces no effect, or one smaller than other interventions, the logic of continuing the program should be called into question given that there are more effective options available. Second, the statistical significance of an effect does not mean the effect is of practical significance. The goal of early reading interventions is for children to no longer be at risk of reading failure at the end of the intervention, or to at least be on this path if it one step in a sequenced structure of support. A significant effect may not be sufficient to alter a student's reading trajectory to remove them from the risk of reading in a timely manner. Several summaries of early reading interventions inform these issues.

The What Works Clearinghouse (WWC) has published numerous *Intervention Reports* based on their evidence review protocol for beginning reading interventions, which is currently in version 2.21 (cite). This review protocol focuses on reading interventions for students in grades K–3 (ages 5–8) that are intended to increase skills in several areas including: alphabetics (phonemic awareness, phonological awareness, letter identification, print awareness, and phonics), reading fluency, comprehension (vocabulary development and reading comprehension), or general reading achievement. The WWC Intervention reports that were categorized by WWC under the categories of “literacy” and “beginning reading” with samples that included students in kinder or first grade are shown in Table 7. Only interventions that included outcomes in alphabetics and general reading achievement are included, since those two domains most closely align to the focus of the OG Institute. Of these sixteen programs, effect sizes for alphabetics were calculated for fourteen and the effect sizes for general reading achievement are listed for five. The effect sizes in alphabetics range from 0.15 to 1.1, with the weighted average being 0.34. The effect sizes in general reading achievement range from 0.22 to 0.61 with a weighted average of 0.29.

Additional to the WWC evidence, Wanzek and Vaughn (2007) reviewed research from 18 studies published between 1995 and 2005. The researchers reviewed interventions that occurred for 100 sessions or longer, which is the equivalent of 20 weeks of daily intervention, similar to what is received in the OG Institute. For the studies reviewed in kindergarten and first grade, Wanzek and Vaughn found effect sized to range from 0.18 to 0.75, with most in the medium effect size range of 0.5. Lastly, the results of NRP found that phonics interventions in kindergarten had an average effect size of 0.56 and first grade had an average effect size of 0.54. The panel suggested that when taught in kindergarten, it must be appropriately designed and must begin with foundational knowledge involving letters and phonemic awareness. Given the range of effect sizes of the aforementioned reviews, an effect size of 0.3 will be considered successful for the intervention.

Conclusion

While randomized controlled trials are considered the gold standard for producing causal estimates of a treatment effect, they are often difficult conduct within public schools. Fien, Smolkowski, Baker, Nelson-Walker & Chaparro (2014) and Baker et al (2015) were able to study a Tier II reading intervention using both an RCT and a RD design. However, both studies were supported by a large grant from the Institute of Educational Sciences. Most districts implement interventions without additional funding and/or additional personnel. In such cases quasi-experimental designs are often necessary to study the impact of new initiatives. In the case of this study, the two methods used can both provide unbiased estimates of the treatment effects of the Orton-Gillingham based reading intervention. Either of these two analysis surpasses previously published research on the Orton-Gillingham Method. The incorporation of both in a single study with strengthen any findings on the efficacy of the program.

CHAPTER 4 RESULTS

Introduction

This chapter will discuss the results of the regression discontinuity and propensity score analysis employed in this investigation. First, the procedures used to verify and clean the dataset will be outlined. Second, a power analysis will be conducted. Third, the dataset will be described and the handling of missing data will be addressed. Fourth, the steps used in the regression discontinuity analysis will be presented, including both the parametric and non-parametric analysis. Fifth, an investigation of the variability of treatment effects will be included. Since students were clustered within schools and classrooms, a multilevel model was used to investigate if the effectiveness of the program varied by school or teacher. Sixth, the details of the propensity score analysis will be discussed. Last, the long-term outcomes of students in the OG Institute will be examined.

Input Errors, Outliers & Normality

The analysis sample was drawn from a large school district in California. Students were selected from schools participating in the OG Institute during their first year of the program. The available data set included 1,801 kindergarten students and 1,702 first grade students who were enrolled in the district during the first administration of DIBELS and had at least one additional DIBELS score recorded during the year. Students who were enrolled in Special Day Classes (SDC) and combination classes were removed from the analysis. One first grade teacher was removed from the analysis because none of the DIBELS scores from the end-of-the-year were entered into the district data system. The remaining cases included 1,655 kindergarten students, with 358 participating in the OG Institute and 1,295 serving as control cases. The first grade sample included 1,458 students with 295 participating in the OG Institute and 1,162 serving as control cases.

Data were screened for possible input errors, outliers and normality. Two kindergarten students and one first grade student had scores higher than the possible score, likely caused by data-entry issues. These were removed from analysis since the original paper scoring sheets were unavailable. Scores that were beyond three standard deviations of the mean were identified as potential outliers, which is a common rule of thumb (Stevens, 2002). A small number of cases in kindergarten were identified as outliers for beginning-of-the-year (BOY) composite scores (n = 15) and end-of-the-year (EOY) composite scores (n = 10). A similar number were also identified in first grade for BOY and EOY composite scores. A 5% trimmed mean was also calculated to find out how much influence was associated with these outlying cases. The trimmed mean on BOY and EOY composite scores did not significantly differ from the original mean, suggesting that outliers did not have much influence on the mean. The univariate outliers were therefore kept in the dataset. Additionally, percentiles for DIBELS composite scores based on national norms were examined (Dewey & Good, 2013). The number of students in the top percentile of DIBELS composite scores is what would be expected given the size of the analytic sample.

Power Analysis

Based on the review of other successful early literacy programs (Table 7), an effect size of 0.3 will be considered the potential outcome for the present investigation. The Optimal Design Software (Spybrook, Raudenbush, Liu, Congdon, & Martínez, 2011) was used to calculate the sample size needed to detect an effect size of 0.3 with a power of 0.80. For a randomized controlled trial (RCT), a sample size of 354 students is needed to detect such a size. However, the statistical power for RD designs is much less than of RCTs and requires a larger sample (Jacob & Zhu, 2012, Lee & Lemieux, 2010). Jacob & Zhu (2012) provide estimates for the necessary increase in sample size required by RD designs. This increase is based on the complexity of the model, the ratio of treatment to control cases, model complexity, and whether or not covariates are used. For a cubic model, the authors estimate a sample size of 3.52 times that of an RCT is necessary for an unbalanced design, where the ratio of treatment to control cases is 3 to 1. About 25% of students were placed in the OG Institute, which is similar to this ratio. The samples used in the following analysis satisfy these assumptions. The kindergarten sample was 4.7 times the size of the RCT and the first grade sample was 4.1 times the sample size needed

Assessing Normality

Histograms of BOY and EOY composites were first visually inspected for both kindergarten and first grade data. Visual inspection can usually suffice to assess the normality of a data set (Warner, 2005). The distribution of kindergarten BOY composite scores was right-skewed and a floor effect was evident. Floor effects occur when tests are too difficult, resulting in some students unable to answer any items correctly (Wang, Zhang, McArdle, & Salthouse, 2009). These two phenomenon were caused by large differences in the entry level literacy skills of the incoming kindergarteners. Many students (n = 121) scored zero on the BOY composite score, which means they were unable to name a single letter or first sound in any of the target words. Several more students had low composite scores, answering only one or two questions correctly. At the other end of the distribution, several students (n = 10) scored in the 99th percentile according to published national DIBELS norms (Dewey & Good, 2013). The histogram for kindergarten BOY scores can be seen in Figure 5.

Floor effects can cause biased estimates in common statistical methods such as ordinary least-squares regression (McBee, 2010). As a result, the outcomes of the parametric RD analysis will be interpreted with reservations. The non-parametric analysis and propensity score analyses, outlined later in this chapter, will be used to assess the potential influence of the floor effect on the RD analysis estimates. The EOY composite scores were roughly normal based on visual inspection for Kindergarten (Figure 6). Other studies have shown floor effects are evident in beginning-of-year administrations of DIBELS but diminish when the assessment is administered again later in the year (Catts, Petsher, Schatschnieder, Bridges & Mendoza, 2008).

First grade BOY composite scores had a normal distribution based on visual inspection, as can be seen in Figure 7. However, first grade EOY composite scores show a bimodal distribution where a larger population of students had low EOY scores compared to BOY scores (Figure 8). This could possibly be caused by the change in tasks required in the EOY assessments. The EOY assessment includes reading connected text – a more complex task not assessed in the BOY assessment. Similar to the kindergarten analysis, the non-parametric RD and propensity score analysis results will be compared to the results of the parametric analysis to investigate if the non-normality possibly affected the parametric modeling of the data.

Missing Data

Missing data is an issue for data analysis (Little & Rubin, 2014). If data are not missing completely at random (MCAR), the results of the analysis may be biased and misleading. No demographic data was missing from the data set. California had implemented a statewide longitudinal database in the time frame of this study and districts were required to regularly submit complete data to the state system. For kindergarten, 0.4% of BOY composite scores were missing and 7.7% of EOY scores were missing, for a total of 8.1% missing values. For first grade, 0.7% of BOY composite scores were missing and 6.7% of EOY scores were missing, for a total of 7.4% missing values. It is considered acceptable to impute the dependent variable (Young & Johnson, 2010). While no consensus has been reached on what amount of missing data is valid, this is considered an acceptable range for imputation (Schlomer, Bauman & Card, 2010). To assess if there are patterns in the missing data, Little's MCAR Chi-square test was used. Little's MCAR test was significant for both kindergarten and first grade datasets, suggesting that the data was not missing at random. Kindergarten's Little's MCAR was $\chi^2 = 169.94$, $p < 0.00$. First grade Little's MCAR was $\chi^2 = 360.6$, $p < 0.00$. The missing values were imputed using the expectation maximization (EM) algorithm. The EM method provides unbiased and efficient parameters (Graham, 2009). To assess if EM was a viable option, a simulation was conducted in which 7% of the EOY scores were randomly deleted and missing values were imputed using EM. The correlation between the imputed values and the actual values was 0.98. Therefore, EM was considered an acceptable approach for replacing the missing data in the present study.

Descriptive Statistics

The demographic information for the kindergarten and first grade samples can be seen in Tables 8 and 9, respectively. Students in the OG Institute were more likely to be poor, male, and English Learners (EL). The BOY composite scores are lower for students in the OG Institute by definition, since scoring below the threshold for risk was used as the criteria to assign students to the intervention. The average scores and standard deviations for BOY and EOY composite scores for kindergarten and first grade can be seen in Table 3.

Analysis 1 - Regression Discontinuity

Assessing Manipulation of the Cut-point

The McCrary Test for discontinuity at the cut-point was conducted to provide evidence that manipulation of assignment to treatment did not occur (McCrary, 2008). Regression equations were created separately on both sides of the cut-point based on histogram-heights of the rating variable. The log difference in the two regression equations' heights at the cut-point was tested to assess if it was statistically significant from zero. In both kindergarten and first grade, the results were not significant. We can conclude that there was not a jump at the cut-point of the rating variable, providing evidence that assignment to the OG Institute was not manipulated by the classroom teacher.

Graphical Representations of the Relationship between Rating and Outcome

A scatterplot of the rating variable (BOY) and outcome variable (EOY) can be seen in Figures 9 and 11. The large number of points makes it difficult to ascertain the relationship of the two variables, or to see if any jump is visible. Lee and Zhu (2009) suggest creating a scatterplot by binning the observations. The bins are single points created from multiple observations, generated by taking the midpoint of the rating variable and the average of the outcome variable. A bin size of four observations was used for kindergarten and a bin size of eight observations was used for first grade. The binned scatter plots can be seen in Figures 10 and 12. A visual inspection of the scatter plots reveals that a discontinuity at the cut-point is not readily evident. Additionally, the binned scatterplots for both kindergarten and first grade appear to be linear around the region of the cut-point, but the linear pattern deviates slightly for observations farther away from the cut-point.

Parametric RD Analysis

Assessing Functional Forms

Specifying the correct functional relationship between the rating variable (DIBELS BOY composite scores) and the outcome variable (DIBELS EOY composite scores) is crucial to estimate an unbiased treatment effect in RD designs (Jacob & Zhu, 2012; Lee & Lemieux, 2010). To assess the functional relationship between the rating and outcome variable, a sample was created from schools within the district that did not implement the OG Institute. Unlike most RD designs, this study was able to utilize a different sample of students to investigate the functional relationship between the rating variable (BOY composite scores) and the outcome variable (EOY composite scores). The schools and students were similar geographically and demographically to the OG Institute schools, making this a viable option. This technique was employed by Baker et al. (2015) in their investigation of a first grade tier II reading intervention. Data was pooled across three years of observations from the schools not implementing OG. The data was fitted using linear, quadratic, and cubic models.

In kindergarten, the model with only a linear term accounted for 32.7% (adjusted R^2) of the variation in the total score. The model with linear and quadratic terms slightly improved the model fit, accounting for 33.0% of the posttest variation. Adding a cubic term improved the prediction, increasing the adjusted R^2 to 33.4%. In addition to the adjusted R^2, the AIC and BIC were also examined (Burnham & Anderson, 2002). Improvements were found with the higher order variables. Similar results were found for the first grade sample. The model with only a linear term accounted for 57.0% (adjusted R^2) of the variation in the total score. The model with linear and quadratic terms slightly improved the model fit, accounting for 57.6% of the posttest variation. Adding a cubic term improved the prediction, increasing the adjusted R^2 to 57.8%. The AIC and BIC also showed improvements for the higher order terms, so the cubic model was used in the parametric analysis. The adjusted R^2, AIC and BIC for the three models can be seen in Table 11. Figures 13 and 14 show the parametric models fitted to the sample from sample of non-OG Institute schools.

Outcomes of Parametric RD Analysis

Kindergarten. The results of the parametric RD analysis show a small but non-significant treatment effect of 8.79 composite score points ($p > 0.2894$). This translates into an effect size of d = 0.27. Figure 15 shows the fitted cubic function on the binned scatterplot of the BOY and EOY composite scores. Although the functional form was determined to be cubic from the analysis on non-OG Institute schools, the linear and quadratic functions were also tested. Neither produced a significant treatment effect, and the size of both effects were smaller than the effect seen from the cubic model.

First Grade. The results of the parametric RD analysis show a small but non-significant treatment effect of 3.18 composite score points ($p > 0.9984$). This translates into an effect size of 0.06. Figure 16 shows the fitted cubic function on the binned scatterplot of the BOY and EOY composite scores. Although the functional form was determined to be cubic from the analysis on non-OG Institute schools, the linear and quadratic functions were also tested. Neither produced a significant treatment effect, and the size of both effects were similar in size to the effect found in the cubic model.

Non-Parametric RD Analysis

The non-parametric regression discontinuity analysis was conducted by performing a local linear regression on both sides of the discontinuity on a subsection of the data. Choosing the optimal bandwidth can be performed by using a formula developed by Imbens and Kalyanaraman (2009):

$$\hat{h}_{opt} = C_K \cdot \left(\frac{2 \cdot \hat{\sigma}^2(c)/\hat{f}(c)}{(\hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c))^2 + (\hat{r}_+ + \hat{r}_-)}\right)^{1/5} \cdot N^{-1/5}$$

where $\hat{h}_{opt}$ is the optimal bandwidth, C_k is a weighting function (in this case a triangular kernel), c is the cut-point value, $\hat{\sigma}^2(c)$ is the estimated conditional variance function of the rating variable at the cut-point, $\hat{f}^2(c)$ is the estimated density function of the rating variable at the cut-point, $\hat{m}_+^2(c)$ and $\hat{m}_-^2(c)$ is the second derivative of the relationship between the outcome and rating variable, r_+^2 and r_-^2 is the regularization term to the denominator in the equation to adjust for the potential low precision in estimating the second derivative, and N is the number of observations (Jacob & Zhu, 2012). The bandwidth was calculated in the R statistical package using the ImbensK function from the RDD package.

The results of the Imbens and Kalyanaraman (2009) formula determined that the optimal bandwidth to perform local-linear regression for kindergarten is 12.48. The cut-point used to assign kindergarteners to the OG Institute was 13, therefore the local linear regression includes all OG Institute students with non-OG Institute students scoring less than 26 on the BOY composite score. Similar to the parametric results, the non-parametric analysis shows a small, non-significant treatment effect of 6.7 (d = 0.21) composite score points ($p > 0.3342$). Figure 17 shows the results of the non-parametric analysis for kindergarten, where the two local linear regression lines are fitted to the binned data.

The optimal bandwidth for first grade was calculated to be 48.42. A composite score of less than 97 was used to assign students to the OG Institute in first grade. Therefore, the sample used in the local linear regression for first grade includes the subset of students in the OG Institute who are closer to the cut-point. Similar to the parametric results, the non-parametric analysis shows a small, non-significant treatment effect of 2.3 (d = 0.04) composite score points ($p > 0.4151$). Figure 18 shows the results of the non-parametric analysis for first grade, where both local linear regression lines are fitted to the binned data.

Heterogeneity of Treatment Effects

The analysis of the overall analytical sample did not produce a treatment effect for kindergarten or first grade. An additional analysis was conducted to investigate if treatment effects varied between schools or teachers. Since students were clustered within schools and classrooms, the heterogeneity of treatment effects was investigated by creating multilevel models that nested the intercept and treatment indicators within schools or classrooms. Heterogeneity of the treatment effects would be evident if a statistically significant estimate of the variance is found for treatment between classrooms or schools. A two-level model (see Equation Set 1 from Chapter 3) was created using the HLM software (Bryk, Raudenbush & Congdon, 2010).

First, a model was created where students were clustered within schools. Kindergarten results were similar to the RD analyses where the treatment effect of participating in the OG Institute was not significant (0.69, $p > 0.825$, d = 0.02). However, the variation in treatment effects across schools was significant ($p < 0.001$). This indicates that the treatment effect varied substantially by school and was likely significant in certain schools. This analysis was repeated with teachers as the level of clustering where the variability of the treatment effects was estimated across teachers without regard to school membership. Similarly, the treatment effect was not significant (0.65, $p > 0.793$, d = 0.02) but the variation in treatment effect was significant ($p < 0.001$). For first grade, the results were almost identical. The school level model produced a non-significant treatment effect (4.77, $p > 0.531$, d = 0.08), but a statistically significant variation in treatment effects was found ($p < 0.001$). At the teacher level, the treatment effect was 7.38 ($p > 0.171$, d = 0.13) and the variation in treatment effect was significant ($p < 0.001$). The differences in treatment effect compared to the RD and propensity score analysis are likely caused by differences in estimation methods used in HLM.

The above results indicate that variations existed in treatment effects between classrooms and schools. To gauge how many schools and teachers produced an effect size of 0.3 or higher, 95% confidence intervals were created around each school's and teacher's slope indicator. A composite score gain 10 points would produce the desired effect size in kindergarten and 17 points were produce the effect size in first grade. Three schools produced such results in kindergarten, which is approximately 16% of the schools. Fourteen teachers, or about 22%, produced gains above that threshold. In first grade, only two schools (13%) and seven teachers (15%) produced effect sizes of 0.3 or higher. The results of the HLM analysis can be seen in Tables 12 – 23.

School and teacher level covariates were added to the model to investigate if they reduced the variation of treatment effects. The proportion of English Learners, Special Education Students, and students qualifying for the National School Lunch Program (NSLP) were added to the level 2 slope equations for the OG Institute. The level 2 covariates where entered into the model individually and then collectively. In both kindergarten and first grade, the variation in treatment effects was still significant after adding the level 2 covariates. Additionally, the deviance statistic was not reduced substantially with the addition of these level 2 covariates. This indicates that school or teacher level characteristics did not explain the variation in treatment effects.

Analysis 2 - Propensity Scores

The second analysis utilized a Nonequivalent Comparison Group Design (NECGD) where propensity scores were used to match treatment and control subjects. Since assignment to the OG Institute was based on a baseline test score within a classroom, significant differences in reading ability existed between the control and treatment subjects within classrooms of the schools implementing the OG Institute. As a result, only a small number of students around the cut-point can be considered similar in reading ability. To increase the number of possible comparison subjects in the propensity score analysis, students from schools who did not implement the OG Institute in any year were pooled to create a sample of potential control candidates. The floor effect that is present in kindergarten BOY composite scores may have introduced bias into the results of the RD parametric analysis (See figure 5). This second analysis provides an additional level of causal inference to the investigation. For kindergarten, this sample included 4,196 students and for first grade the sample included 3,747 students. To perform the propensity score matching, the probability of being assigned to the treatment was first calculated using a probit model. In the probit model, the treatment condition is the dependent variable and student covariates, including the BOY DIBELS score, were used as the independent variables. Second, the propensity scores were used to create a comparison group and the balance between the control and treatment groups was ascertained. Third, the treatment effect was calculated.

Creating Propensity Scores

Propensity scores were created using a probit model with treatment membership as the dichotomous outcome variable. The variables from Table 6 were used as covariates in the probit model. The model for both kindergarten and first grade was significant at the $p < 0.01$ level. The model specifications are of secondary importance to the resulting balance between treatment and control groups (Cook & Steiner, 2010; Shadish & Steiner, 2010). Meaning, the size of the coefficients in the probit model relative to each other are not meant to be interpreted as theoretically important. It is the resulting similarity in composition between the control and treatment groups that is critical. To assess the balance between treatment and control groups, as well as calculate the treatment effect, five strata were created within the range of common support between treatment and control subjects.

The range of common support was found by taking the minimum and maximum propensity scores for the treatment cases. For kindergarten, the propensity scores ranged from 0.038 to 0.470 and for first grade, the propensity scores ranged from 0.2 to 0.75. Observations from the control subjects within this range of propensity scores were used to create a comparison group. For kindergarten, the treatment group contained 358 students and the control group contained 1,407 students. For first grade, the treatment group contained 289 students and the control group contained 1,812 students. Within this range of common support, quintiles were used to create strata.

Assessing Balance

Prior to matching, the control subjects and treatment subjects differed substantially on most covariates. Tables 24 and 25 show the summary of covariates in the treatment and control groups before and after matching for kindergarten and first grade, respectively. Several statistical analyses were performed to assess the balance between the resulting control and treatment groups. For the categorical variables, which included all of the covariates other than the BOY composite scores, differences in population proportions were tested for significance for the overall sample as well as within each stratum.

Over the entire sample, the tests showed non-significant differences in the population proportions between treatment and control groups except in English learners and students qualifying for the NSLP. In both cases, the proportions were higher in the treatment group, which may cause an attenuation in the treatment effect. Within each stratum, the tests showed non-significant differences in the population proportions between treatment and control groups. Within-stratum comparisons could not be made for Asian students, African American students, and students of other ethnicities because of the small number of subjects with those characteristics in these stratums.

For the single continuous covariate, BOY composite scores, kernel density plots were created before and after matching. The results can be seen in Figures 19 and 20. After matching, the distribution of BOY composite scores between treatment and control subjects are much more similar. Additionally, the ratio of variances on BOY composite scores was calculated between treatment and control subjects, which can be a further test of balance (Shadish & Steiner, 2010). Balance is met if the ratio of variances is between 4/5 and 5/4 (Cook & Steiner, 2010). The variance ratio was within this range for kindergarten (TREAT s^2 / CONTROL $s^2 = 1.12$) and first grade (TREAT s^2 / CONTROL $s^2 = 1.18$).

Calculating the Treatment Effect

The treatment effect was calculated by taking the weighted averages of the differences in outcomes between each of the five strata. The range of propensity scores for each of these and the average composite scores for treatment and control subjects can be seen in Tables 26 and 27. The kindergarten analysis produced a treatment effect of 3.8 composite score points, an effect size of d = 0.12. The first grade analysis produced a treatment effect of 4.6 composite score points. This is equivalent to an effect size of 0.08. Since students were clustered within schools and classrooms, the standard errors were adjusted (Cameron & Miller, 2013). The intraclass correlation coefficient was calculated for both clustering at the school and classroom level. Based on these adjusted standard errors, neither the results for neither kindergarten or first grade were statistically significant.

Short and Long Term Outcomes of OG Institute Students

The risk categories for EOY DIBELS composite scores were also used to assess the short- and long-term outcomes of students that participated in the OG Institute. At the beginning of the year, all students placed in the intervention were categorized as *well-*

below benchmark. Students in this group have a 10 to 20% chance of meeting later DIBELS benchmarks without intensive intervention (Good et al, 2011). The authors of DIBELS consider students in the *below benchmark* category to need intervention, but of less intensity than those in the *well-below benchmark* category. Students in this group have a 40 – 60% chance of making later reading targets without additional support. Students who meet the benchmark targets are considered *at benchmark*. The goal of a reading intervention is to have all students at eventually succeeding in the core curriculum. Students in the *benchmark* category are considered to be progressing appropriately with the core curriculum (Good et al, 2011).

Short-Term Results

After being enrolled in the OG Institute for one year, 115 of the 358 (32%) kindergarten students were meeting benchmark targets. An additional 88 (25%) students, improved to the *below benchmark* category. There was still a substantial proportion, 155 (43%), of students in the *well below benchmark* category. For first grade, 93 of the 295 (32%) students were meeting grade level targets. An additional 12% (n = 36) were *below benchmark*, but over half (56%, n = 166) were still in the *well-below benchmark* category. These results can be seen in Tables 28 and 29.

Long Term Results

Since the OG Institute was initiated to address the ‘Reading by Third’ goal of the district, EOY scores from second grade were used as the long term indicator of reading by third grade. For students in the two earliest cohorts (2011-12 and 2012-13 school years), second grade EOY data were available. For kindergarten, scores were available for 135 students that participated in the OG Institute. For first grade scores for 152 students were available. At the end of second grade, 28% of kindergarten students who participated in the OG Institute were classified as *at benchmark*. For first grade, 17% were classified as *at benchmark* at the end of second grade.

CHAPTER 5 DISCUSSION

Introduction

This chapter will discuss the results from the regression discontinuity and propensity score analysis used to investigate the effectiveness of the OG Institute. First, the summary of key findings will be presented. These findings will be contrasted to results of previous Orton-Gillingham based reading interventions as well as other successful early literacy interventions. The limitations of the current work will then be presented. A discussion of the implications of this study will follow. Lastly, directions for future research will be suggested.

Summary of Key Findings

Treatment Effects from the RD Analysis

For kindergarten students participating in the OG Institute, the results of both the parametric and non-parametric regression discontinuity analysis showed small, but non-significant improvements to their end-of-year reading scores. The parametric RD analysis produced a treatment effect of 8.79 EOY DIBELS composite score points, an effect size of 0.27. However, the result was not statistically significant (p = 0.29). In a parametric RD analysis, the model chosen to represent the relationship between the rating and outcome variables is critical. Misspecification of the functional form can influence the magnitude of the treatment effect at the cut-point (Jacob & Zhu, 2012). Reasonable measures were taken to correctly model the functional form based on previous studies and best practices (Baker et al, 2015, Jacob & Zhu, 2012, Lee & Lemieux, 2010). Additionally, all parametric models that should be considered during an RD analysis

(Jacob & Zhu, 2012) did not produce significant results, providing further evidence that the model chosen was not the cause of the non-significant effect.

The results of the analysis using a cubic model did approach an effect size similar to other successful reading interventions, but the floor effect evident in the kindergarten BOY composite scores may have biased these results. This, coupled with the results of the non-parametric analysis, suggests that the treatment effect for kindergarten is below the 0.3 effect size threshold chosen as the indicator of success for the intervention.

Similar to kindergarten, the results of both the RD parametric analysis and non-parametric analysis for the first grade program showed small, non-significant treatment effects. The effect size of the parametric model was 0.05 and the effect size of the non-parametric model was 0.04. Changes to the parametric model did not produce significant results, which further substantiates that a large effect was not present. While the results of the parametric analysis showed a slightly larger effect than the non-parametric results, neither approached the size that could be considered a substantial improvement in reading scores compared to those students receiving only core classroom instruction.

Treatment Effects from Propensity Score Analysis

The propensity score analyses found a small effect in both kindergarten and first grade. An effect in of 3.8 composite score points was found for the kindergarten program and an effect of 4.6 composite score points was found for the first grade program. This translates into an effect size of 0.12 and 0.08 respectively, which is the threshold of a small effect (Cohen, 1992). However, when adjusting the standard errors for clustering, the results were also not significant.

The Variability of Treatment Effects

The heterogeneity of treatment effects was investigated using multilevel modeling. This analysis also found no treatment effect, but the effects varied significantly between schools and teachers. In kindergarten, the treatment effect was 0.69 ($p > 0.85$) and the effects varied significantly between schools and classrooms. Only three schools produced effects in the range of other successful early literacy programs with outcomes above 10 composite score points ($d > 0.30$). The effect found when students were clustered at the teacher level was 0.65 with a ($p > 0.79$). Fourteen kindergarten teachers produced effect sizes higher than 0.30.

The first grade results were very similar. The school level analysis showed an effect of 4.8 composite score points ($p > 0.51$). Only two schools produced an effect of more than 17 composite score points, which is greater than an effect size of $d > 0.30$. The teacher level analysis showed an overall treatment effect of 7.4 ($p > 0.17$). Seven first grade teachers produced an effect size of 0.3.

The composition of schools and classrooms did not explain a substantial amount of this variation. When the proportion of English Learners, Student with Disabilities, and student receiving Free and Reduced lunch were added as school or teacher level covariates, the variation did not significantly decrease. Other unmeasured school or classroom characteristics, such as teacher or school effectiveness, are potential causes of this variation.

Comparison of Treatment Effects

The treatment effects found in the three analyses were similar but varied somewhat. For kindergarten the effect of OG in terms of EOY composite scores varied from 0.65 to 8.79. The first grade effects ranged from 2.3 to 7.4. This translates into an effect size range of 0.02 to 0.27 for kindergarten and 0.04 to 0.13 for first grade. The effect size of the kindergarten parametric analysis produced the largest effect size of 0.27. However, the floor effects present may have biased these results. Furthermore, the other analysis did not approach this size and no result was statistically significant. A few reasons may be causing their slight variation in magnitudes. First, the analyses used different comparison groups. The RD analysis compared students within classrooms of schools implementing the OG Institute. The propensity score analysis also included students in schools that did not participate in the OG Institute. Second, each analysis used different estimation methods. Although the effects varied slightly, all were small, not significant, and less than the effects found in other successful early reading interventions.

While the power analysis showed that there was a sufficient sample to detect an effect size of 0.3 but the actual effects of the program were much smaller. A sample of about twice the size of the present study would be necessary to detect an effect in the 0.1 range. However, increasing the sample size to detect a smaller effect would not be productive, considering that alternative programs exist that have produced substantially larger effects (Table 7). Furthermore, the improvements to EOY composite scores for students participating in the program do not translate into practical significance in their reading scores. The increases to EOY composite score found in in this study represent about a three-percentile increase. This is not enough to get most at-risk students across the bar into reading proficiency. The results show that the OG Institute provided marginal improvements at best. A comparison of these outcomes to previous OG studies and other early literacy initiatives follows.

Findings in Context of Previous Literature

Outcomes of Previous OG Based Reading Interventions

As outlined in the literature review, few previous studies on Orton-Gillingham based reading interventions have been designed to provide a causal estimate of a treatment effect. The review of Orton-Gillingham based interventions by Ritchey and Goeke (2006) showed that a few early elementary examples of OG based interventions produced successful results. These studies (Litcher & Roberge, 1979; Stoner, 1991; Joshi, Dalhgren & Boulware-Gooden, 2002) had large effects of more than 0.75. However, because of their shortcomings and no corroborating evidence from well-designed studies, the results should be taken with reservations. Two Masters' Theses on Project Read met the minimum requirements of the WWC. The first by Acalin (1995) studied kindergarten through fourth grade students in Resource Specialist Program (RSP) classrooms. Two groups were given OG-based reading interventions. While both improved, no differences between the groups were found. The second, conducted by Bussjaeger (1993) included fourth and fifth grade students and found Project Read to be more effective than literature-based instruction used in the control classrooms.

Effect Sizes of non-OG Based Programs

Many earlier literacy programs not based on the Orton-Gillingham method have produced positive results. Based on a review of What Works Clearinghouse (WWC) *Intervention Reports* several early reading programs produced significant results in alphabetics (phonemic awareness, phonological awareness, letter identification, print awareness, and phonics), reading fluency, comprehension (vocabulary development and reading comprehension), or general reading achievement. The effect sizes in alphabetics ranged from 0.15 to 1.1, with the weighted average being 0.34. The effect sizes in general reading achievement ranged from 0.22 to 0.61 with a weighted average of 0.29.

Additionally, Wanzek and Vaughn (2007) reviewed research from 18 studies published between 1995 and 2005 with similar intensity to the OG Institute. The studies that took place in kindergarten and first grade produced effect sizes from 0.18 to 0.75, with most in the medium effect size range of 0.5. The results of the National Reading Panel report (2000) found that phonics interventions in kindergarten had an average effect size of 0.56 and first grade had an average effect size of 0.54. Lastly, Baker et al (2015) found students receiving both Tier 1 instruction and Tier 2 intervention achieved significantly higher reading scores at the cut point compared to students who received Tier 1 only. The researchers found an effect of approximately 8 percentile points.

Given the range of effect sizes of the aforementioned reviews, an effect size of 0.3 was considered the threshold for success of the OG Institute. This effect size would have translated into a 10-point difference in EOY composite scores for kindergarten and 17-point EOY composite score difference in first grade. Given that most results for both grades resulted in gains of about half this size, it is evident that the results of the OG Institute fell short of other successful early literacy interventions.

Effect on Risk Status

Several previous studies tracked the short- and long-term reading outcomes for students receiving reading interventions. The results of this work show that most students identified at-risk in early grades respond well to structured interventions. Lyon et al (2001) noted that 85-90% of poor readers can increase reading skills to average reading levels when provided timely, well-structured intervention. Similarly, Torgesen (2000) showed that the majority of students placed in early reading interventions improved their reading achievement. His estimates showed that early intervention can reduce the number of children who develop reading difficulties by up to 70%. Additionally, Vellutino et al. (2006) found that 84% of students who received remediation in kindergarten were no longer at risk by the end of third grade.

The outcomes of this study produced much smaller success rates. Many of the kindergarten students who participated in the OG Institute were still struggling in reading at the end of the first year. About one-third of them met the reading targets as defined by DIBELS at the end of the year. While lower than the aforementioned studies, this is better than the prediction made by the defined risk categories in DIBELS (Good et al, 2010), which states a lower outlook of 10-20%. Another 25% improved into the *below benchmark* DIBELS category. However, 43% were still at *well below benchmark* at the end of one year of treatment. At the end of second grade, the results diminished slightly;

28% of students who had been enrolled in the OG Institute were meeting grade level reading targets.

Smaller success rates existed for first grade. Many of the students who participated in the OG Institute were still struggling in reading at the end of the year. About 32% of those in the OG Institute were meeting grade level reading targets at the end of first grade. An additional 12% improved their classification to *below benchmark*. However, over half (56%), were still at the well below benchmark level. At the end of second grade, 24% of the students that participated in the OG Institute were reading at grade level as measured by DIBELS.

Limitations

Although this study is the first that meets the WWC design criteria to investigate an Orton-Gillingham based early literacy intervention, several limitations should be noted. First, additional fidelity checks may have improved the program implementation. Fidelity checks were conducted during each school's initial year of implementing the OG Institute; however, they may have been insufficient. Only two instructional staff supported all seventeen schools. Having additional expert reading staff oversee and work with teachers on a more regular basis may produce better outcomes for the intervention. For example, Vadasy et al (2005) provided weekly on-site coaching and modeling during their year-long study of an early reading intervention. The researchers also used this time at school sites to gather observation data on treatment fidelity. The effect size of this study was 0.91, substantially larger than the OG-Institute.

Second, the training provided to teachers may have been insufficient. Teachers in the OG Institute participated in a week-long, forty-hour workshop, similar to what was provided by other successful programs (Blachman et al, 1999). However, additional training may have been needed based on the skill level of the teachers. There is evidence (Joshi, Binks, Hougen, Dahlgren, Ocker-Dean, Smith, 2009) that many teachers lack the training and background to adequately teach early literacy skills. No needs assessment was conducted to ascertain what amount of training was necessary. Also, no evaluations of the training were conducted. Furthermore, teachers were not assessed on whether or not they learned the instructional strategies and theories behind the program. Other large scale PD implementations monitored participant engagement during the training as well as the pre- and post- content knowledge of the participants (Garet, Wayne, Stancavage, Taylor, Walters, Song & Doolittle, 2010).

Third, floor effects were found for BOY DIBELS scores which were used to assign students to the OG Institute. Catts et al (2008) discovered that beginning-of-year administrations of DIBELS produced floor effects resulting in poor predictability of which students would benefit from tier II interventions. Furthermore, floor effects have been found to introduce bias into estimates (McBee, 2010; Wang et al, 2009). Subsequently, the kindergarten RD results may have been biased.

Fourth, additional covariates may have been beneficial in the propensity score analysis. In propensity score analysis, the assumption of strong ignorability must be met (Rosenbaum & Rubin,1983). While there is no empirical test of strong ignorability (Austin, 2011), all covariates that may influence the outcome of the treatment should be included in the calculation of propensity scores. While this study included many common covariates associated with achievement, such as English Learner status, Special Education services, ethnicity and socio-economic background, they may have been insufficient. In particular, pre-test measures are important to include (Cook & Steiner, 2010). However, as noted by Catts et al (2008), BOY DIBELS scores alone may not provide adequate predictability in identifying which students would benefit from reading interventions. Including additional early reading measures may have improved the balance between treatment and control groups.

Implications for the Use of OG

The OG Institute was initiated by the district as the main component of their 'Reading by Third' initiative. While many districts do not explicitly state a reading goal in this manner, having students develop proficient reading skills in a timely manner is a fundamental goal of all schools. The consequences of students not reaching reading proficiency within the normally developed timeline have been well documented (Kutner, Greenberg, Jin, Boyle, Hsu & Dunleavy, 2007; Lloyd, 1978; Reschly, 2010; Donovan & Cross, 2002). Based on both the short- and long- term outcomes of students in the OG Institute, the program did not provide the expected results. As such, districts should be cautious when contemplating the use of an OG-based reading intervention in general education settings.

A few reasons may have caused the poor results. Theory failure (Rossi, Lipsey, & Freeman, 2004) refers to the scenario that the links between intended changes in instruction and intended student outcomes have proven incorrect. The instructional composition of the program may not have been properly designed. For example, the proportion of instruction spent on varying early literacy skills may not have been well balanced. Additionally, the multisensory instruction (MSI) components may not be effective. Despite the current support of multisensory instruction by organizations and practitioners, sparse empirical evidence exists for MSI (Uhry & Clark , 1995; Birsch, 2005). The merits of including MSI in a reading program are still debated (Spear-Swerling, 2013). In studies that have investigated multisensory techniques, none have isolated the influence of the multisensory methods (Uhry & Clark, 2005; Fletcher, Lyon, Fuchs, & Barnes, 2006). Wise, Ring, and Olson (1999) did not find that the multisensory articulatory component of the Lindamood program was necessary when they compared it to the results that did not include it. No studies have yet been designed to isolate the impact of MSI on literacy instruction (Spear-Swerling, 2013).

Few high quality studies on OG-based interventions have been previously conducted. One potential explanation is the "publication bias", also known as the "file drawer problem" (Rosenthal, 1979; Scargle, 2000). A result of publication bias is that other studies of OG may have similarly produced non-significant results, but were not published.

[paragraph on how it may just be the organization of this intervention]

Lastly, substantial costs are associated with providing professional development to teachers and for the instructional materials needed for reading interventions. The costs for reading interventions generally increase as students get older (Hollands, Pan, Shand, Cheng, Levin, Belfield, & Hanisch-Cerda, 2013). The financial burdens felt by district to provide remediation for older readers could potentially be avoided or diminished if high quality early intervention programs are chosen.

Directions for Future Work

Although multisensory instruction has gained support of practitioners and professional groups, no studies have yet isolated its benefit in early literacy instruction (Spear-Swerling, 2013). A well-designed randomized controlled trial could help identify if multisensory instruction is a beneficial delivery method for early literacy skills. Classrooms within different schools and district could be assigned to either an OG-based intervention or an intervention without any multisensory components. A purported benefit attributed to MSI is the engagement of the student and resulting focus on sound-symbol relationships. A valid comparison would need to include a mode of instruction that could be arguably similar in engagement. Additionally, both programs would need to follow the same sequence of concepts and devote similar amounts of time to each skill. As mentioned in the previous section, delaying the administration of the screening instrument until later in the kindergarten year may provide a better indication of which students require tier II intervention. However, this will reduce the time for interventions. As suggested by Catts et al, the DIBELS assessment could be used as a “first cut” to identify children who do not meet the benchmark, and a follow-up phonological awareness measure could be used to help reduce the false-positives that are introduced by floor effects.

Conclusion

Previous research has established the importance of children acquiring early literacy skills, with kindergarten and first grade being a crucial period to identify and remediate students who are not on track to read proficiently. While OG has been popular for over half a century and is endorsed by the American Dyslexic Foundation and the International Dyslexic Foundation, there has been sparse evidence proving the efficacy of program. It has been proposed that the widespread use of OG instruction has been driven by anecdotal evidence and personal experience (Ritchey & Goeke, 2006). In the last decade, the lack of methodologically sound studies and need for scientifically based evidence of OG's effectiveness (Alexander & Slinger-Constant, 2004; Rose and Zirkel, 2007) has been identified. This study contributed to the literature on early literacy interventions by using a design that met the WWC criteria for showing causal inference. Additionally, further evidence was provided through a second analysis using a Nonequivalent Comparison Group Design (NECGD) with propensity scores. The results of both analyses revealed no effect for students enrolled in the intervention in either kindergarten or first grade. Long-term outcomes showed that over half of the students in the intervention were still not meeting reading targets by the end of second grade. The multilevel analysis showed that the treatment effect varied significantly across classrooms and schools. However, no available measures of classroom or school characteristics where associated with that variation. Together, these findings suggest, on average, the OG-based intervention reviewed here was ineffective as a general education intervention for struggling early readers. Certain schools and classrooms did exist classrooms where it was effective. This study provides evidence that the popularity of OG may be based on factors other than student success and that the program's migration to general education reading instruction may be unwarranted. Future research is needed to investigate whether specific characteristics of classrooms and schools are associated with the effectiveness of OG, and if so, whether effective characteristics can be replicated to scale.

TABLES

Table 1

Demographic information for each elementary school in the sample.

School	**Cohort**	**Total Enrollment**	**K/1 Enrollment**	**Hispanic**	**White**	**Black/ African Am**	**Other**	**SED**	**EL**
1	Cohort 1	798	234	66%	13%	13%	8%	71%	31%
2	Cohort 1	817	238	52%	18%	17%	13%	76%	31%
3	Cohort 1	790	245	94%	2%	1%	2%	94%	72%
4	Cohort 1	763	221	82%	9%	3%	5%	79%	32%
5	Cohort 1	779	234	70%	21%	4%	5%	71%	29%
6	Cohort 2	609	206	61%	28%	5%	6%	50%	28%
7	Cohort 2	544	173	78%	9%	7%	5%	91%	45%
8	Cohort 2	847	228	78%	11%	6%	5%	84%	46%
9	Cohort 2	671	200	78%	13%	4%	5%	86%	37%
10	Cohort 2	702	196	66%	19%	6%	9%	67%	35%
11	Cohort 2	660	183	38%	48%	5%	9%	44%	9%
12	Cohort 3	773	193	57%	29%	9%	6%	62%	23%
13	Cohort 3	473	131	63%	27%	6%	4%	63%	15%
14	Cohort 3	548	160	64%	22%	8%	6%	65%	24%
15	Cohort 3	793	219	85%	7%	2%	7%	89%	55%
16	Cohort 3	704	194	62%	24%	5%	8%	68%	24%
17	Cohort 3	712	206	79%	6%	8%	7%	82%	31%
18	Cohort 3	580	180	76%	15%	4%	6%	72%	31%
19	Cohort 3	799	211	65%	25%	3%	7%	57%	31%
20	No OG	554	176	75%	15%	4%	6%	79%	42%
21	No OG	838	223	33%	44%	7%	16%	25%	8%
22	No OG	713	240	72%	15%	6%	7%	75%	32%
23	No OG	669	201	89%	7%	1%	3%	85%	46%
24	No OG	910	267	80%	12%	4%	4%	79%	41%
25	No OG	1090	256	33%	42%	6%	19%	39%	8%
26	No OG	907	249	27%	47%	4%	22%	27%	8%
27	No OG	718	222	79%	10%	5%	5%	80%	39%
28	No OG	1093	305	42%	34%	9%	15%	37%	7%
29	No OG	755	215	39%	41%	5%	14%	35%	4%
District Total		**21609**	**6206**	**64%**	**22%**	**6%**	**9%**	**66%**	**30%**

Note: In the Cohort column, No OG refers to the schools that did not implement the OG Institute in any year of this analysis. These schools were used to provide control students in the propensity score analysis and assess the functional form in the regression discontinuity analysis. (SED = socioeconomically disadvantaged, EL = English Learner)

Table 2

Timeline of the implementation of the OG Institute.

Cohort	K1 Enrollment (2014-15)	Number of Schools	2010-11	2011-12	2012-13	2013-14	2014-15
1	1172	5	-	OG	OG	OG	OG
2	1379	7	-	-	OG	OG	OG
3	1301	7	-	-	-	-	OG
Non OG Schools	2354	10	-	-	-	-	-

Note: The five schools in Cohort 1 began implementing the OG Institute in 2011-12 and continued the program in the following years. Schools in Cohort 2 began two years later and schools in cohort 3 began in 2014-15. Enrollment is from the 2014-15 school year but was relatively stable over the course of the five years.

Table 3

The ten lessons, or steps, in the OG Institute lesson weekly plan

Step	**Lesson**
Step 1	Phonemic Awareness
Step 2	Three-Part Drill
Step 3	Vowel Intensive
Step 4	Introducing a New Concept A: Letter Formation
Step 5	Introducing a New Concept B: Sound-Symbol
Step 6	Introducing a New Concept C: Making Connections to Prior Knowledge
Step 7	Spelling and Blending with Letter Tiles
Step 8	Dictation
Step 9	Red Words
Step 10	Reading Fluency

Note: Details of each lesson can be seen in Appendix A. The OG Institute takes place daily for about 20 minutes. The instructor typically incorporates two steps into each day's twenty minute session. The steps are continuously cycled, with specific topics of each step changing with each rotation.

Table 4

DIBELS assessments and their relationship to early literacy skills

Early Literacy Skill	**DIBELS Indicator**
Phonemic Awareness	First Sound Fluency (FSF)
	Phoneme Segmentation Fluency (PSF)
Alphabetic Principle and Basic Phonics	Nonsense Word Fluency (NWF) -Correct Letter Sound -Whole Words Read
Advanced Phonics and Word Attack Skills	DIBELS Oral Reading Fluency (DORF) -Accuracy
Accurate and Fluent Reading of Connected Text	DIBLES Oral Reading Fluency (DORF) -Correct Words Per Minute -Accuracy

Table 5

Timeline for the assessments administered as part of DIBELS

DIBELS Task	Kindergarten			First Grade			Second Grade		
	BOY	*MOY*	*EOY*	*BOY*	*MOY*	*EOY*	*BOY*	*MOY*	*EOY*
FSF	x	x							
LNF	x	x	x	x					
PSF		x	x	x	x	x			
NWF		x	x	x	x	x	x	x	x
DORF					x	x	x	x	x

Table 6

Variables used in the HLM model and for propensity score matching

Variable	Description
OG	Student participated in the OG intervention (0 = not in intervention, 1 = in intervention)
$DIBELS_{BOY}$	The composite score on the beginning of the year DIBELS assessment
NSLP	Student receives free or reduced lunch as part of the National School Lunch Program (0 = not in NSLP, 1 = in NSLP)
HSG	Highest education of either parent is a high school graduate
SOMECOLL	Highest education of either parent is some college
COLLGRAD	Highest education of either parent is a college graduate
FEMALE	Student is female
EL	Student is an English Language Learner
SWD	Student has a learning disability
WHITE	Student is white
ASIAN	Student is Asian
AA	Student is African American
OTHERETHNICITY	Student is a ethnicity other than white, Asian, African American, or Hispanic

Note: A student with 0 for HSG, SOMECOLL, and COLLGRAD has neither parent earning a high school diploma. A student coded with 0 for WHITE, ASIAN, AA, and OTHERETHNICITY is Hispanic, which is the largest ethnic group in the district

Table 7

The effect sizes of interventions in WWC intervention reports focusing alphabetics and general reading achievement between 2008 & 2015

Intervention	**Grades**	**Sample Size**	**Delivery**	**Alphabetics**	**General Reading**
Early Intervention in Reading	1	59	Supplemental	1.1	
Earobics	K-3	246	Computer/ Supplemental	0.49	
Fast ForWord	K-4	1390	Computer/ Supplemental	0.15	
Ladders to Literacy	K	760	Supplemental	0.69	
Lexia	K-1	314	Computer/ Supplemental	0.27	0.23
Lindamood Phonemic Sequencing (LiPS)	1	150	Computer/ Supplemental	0.45	
Little Books	K	325	Supplemental		0.31
Peer-Assisted Learning/Literacy Strategies (PALS)	K-1	3130	Supplemental	0.35	
Read, Write & Type	1	150	Computer/ Supplemental	0.19	
Reading Recovery	1	866	Supplemental		0.61
Sound Partners	K-1	442	Supplemental	0.55	0.22
Start Making a Reader Today (SMART)	1	125	Supplemental	0.4	
Stepping Stones to Literacy	K	120	Supplemental	0.84	
Success for All	K-3	4000	Whole Class	0.33	0.26
Voyager Universal Literacy System	K	600	Whole Class	0.28	
Waterford Early Reading Program	K	70	Computer/ Supplemental	0.5	
		Weighted Average		**0.34**	**0.29**

Note: The weighted average was based on sample size and effect size

Table 8

Demographic Information for Kindergarten

	Kindergarten	
	Not in OG Institute (control)	*In OG Institute (treatment)*
Enrollment	1295	358
Ethnicity		
White	209 (16.1%)	41 (11.4%)
African American	33 (2.5%)	5 (1.4%)
Hispanic	1012 (78.2%)	304 (84.9%)
Asian	14 (1.1%)	2 (0.5%)
Other Ethnicity	26 (2.0%)	6 (1.7%)
Special Education	61 (4.7%)	31 (8.7%)
English Learner	427 (33.9%)	168 (46.9%)
Female	684 (52.8%)	162 (45.2%)
Socioeconomically Disadvantaged	545 (42.1%)	172 (48.0%)

Table 9

Demographic Information for First Grade

	First Grade	
	Not in OG Institute (control)	*In OG Institute (treatment)*
Enrollment	1162	295
Ethnicity		
White	183 (15.7%)	27 (9.2%)
African American	24 (2.1%)	5 (1.7%)
Hispanic	937 (80.6%)	258 (87.5%)
Asian	8 (0.7%)	2 (0.7%)
Other Ethnicity	10 (0.9%)	3 (1.0%)
Special Education	56 (4.8%)	47 (15.9%)
English Learner	605 (32.9%)	131 (44.4%)
Female	684 (52.1%)	121 (41.0%)
Socioeconomically Disadvantaged	514 (39.7%)	142 (48.1%)

Table 10

OG Institute Cohort DIBELS BOY & EOY Scores Kindergarten

	Kindergarten		First Grade	
	Not in OG Institute (control)	*In OG Institute (treatment)*	*Not in OG Institute (control)*	*In OG Institute (treatment)*
BOY	33.2 (22.8)	3.1 (3.5)	124.9 (37.9)	60.3 (28.3)
EOY	131.3 (49.3)	99.9 (41.3)	193.3 (77.0)	106.2 (84.1)

Note: Mean (Standard Deviation)

Table 11

Regression Statistics Used to Assess the Functional Form of the Regression Discontinuity Analysis. The cubic form was chosen for both kindergarten and first grade.

	Kindergarten			First Grade		
	Adj. R^2	*AIC*	*BIC*	*Adj.* R^2	*AIC*	*BIC*
Linear	0.3397	42992.3	43062.05	0.5778	38387.6	38455.31
Quadratic	0.3425	42974.96	43044.72	0.584	38337.12	38410.98
Cubic	0.3472	42946.84	43029.29	0.5857	38324.13	38404.14

Table 12

Variance Components for the HLM Unconditional Kindergarten School Model

Random Effect	Standard Deviation	Variance Component	*d.f.*	χ^2	*p*-value
INTRCPT1, u_0	26.56832	705.87579	18	620.79173	<0.001
level-1, r	42.37384	1795.54240			

Table 13

Variance Components for the HLM Unconditional Kindergarten Teacher Model

Random Effect	Standard Deviation	Variance Component	*d.f.*	χ^2	*p*-value
INTRCPT1, u_0	27.65	764.73	64	756.14	<0.001
level-1, r	41.55	1726.45			

Table 14

Variance Components for the HLM Unconditional First Grade School Model

Random Effect	Standard Deviation	Variance Component	*d.f.*	χ^2	*p*-value
INTRCPT1, u_0	21.50	462.33	15	112.86	<0.001
level-1, *r*	82.55	6816.10			

Table 15

Variance Components for the HLM Unconditional First Grade Teacher Model

Random Effect	Standard Deviation	Variance Component	*d.f.*	χ^2	*p*-value
INTRCPT1, u_0	27.93	780.43	47	221.66	<0.001
level-1, *r*	80.60	6496.46			

Table 16

Fixed Effects for the HLM Kindergarten School Model

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
INTRCPT, γ_{00}	97.21	6.76	14.37	18	<0.001
ISFEMALE, γ_{10}	2.74	1.50	1.82	1606	0.068
ISASIAN, γ_{20}	1.34	12.20	0.11	1606	0.913
ISWHITE, γ_{30}	1.52	2.92	0.52	1606	0.603
ISAA, γ_{40}	6.14	4.86	1.26	1606	0.207
ISOTHER, γ_{50}	-3.08	4.84	-0.63	1606	0.525
ISEL, γ_{60}	-4.27	1.73	-2.46	1606	0.014
ISSWD, γ_{70}	-11.68	3.78	-3.09	1606	0.002
ISNSLP, γ_{80}	-4.26	1.31	-3.24	1606	0.001
DIBELSBEG, γ_{90}	1.04	0.04	21.70	1606	<0.001
INOGINST, γ_{100}	0.69	3.08	0.22	18	0.825

Table 17

Variance Components for the HLM Kindergarten School Model

Random Effect	Standard Deviation	Variance Component	*d.f.*	χ^2	*p*-value
INTRCPT1, u_0	31.67	1003.16	18	1046.26	<0.001
INOGINST slope, u_{10}	11.39	129.77	18	50.10	<0.001
level-1, r	33.83	1144.87			

Table 18

Fixed Effects for the HLM Kindergarten Teacher Model

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
INTRCPT, γ_{00}	99.57	4.58	21.70	64	<0.001
ISFEMALE, γ_{10}	2.51	1.43	1.75	1514	0.079
ISASIAN, γ_{20}	1.52	9.04	0.16	1514	0.866
ISWHITE, γ_{30}	-0.71	2.26	-0.31	1514	0.751
ISAA, γ_{40}	-0.88	6.19	-0.14	1514	0.886
ISOTHER, γ_{50}	-3.98	5.29	-0.75	1514	0.452
ISEL, γ_{60}	-4.91	1.60	-3.05	1514	0.002
ISSWD, γ_{70}	-12.21	3.94	-3.09	1514	0.002
ISNSLP, γ_{80}	-4.94	1.69	-2.92	1514	0.004
DIBELSBEG, γ_{90}	1.04	0.04	23.47	1514	<0.001
INOGINST, γ_{100}	0.64	2.44	0.26	64	0.793

Table 19

Variance Components for the HLM Kindergarten Teacher Model

Random Effect	Standard Deviation	Variance Component	*d.f.*	χ^2	*p*-value
INTRCPT1, u_0	31.93	1020.06	63	1209.09	<0.001
INOGINST slope, u_{10}	10.35	107.25	63	87.03	0.024
level-1, r	32.81	1077.13			

Table 20

Fixed Effects for the HLM First Grade School Model

Fixed Effect	Coefficient	Standard error	t-ratio	Approx. *d.f.*	p-value
INTRCPT, γ_{00}	19.96	7.45	2.67	15	0.017
ISFEMALE, γ_{10}	3.38	3.13	1.07	1416	0.282
ISASIAN, γ_{20}	-25.10	8.46	-2.96	1416	0.003
ISWHITE, γ_{30}	2.10	3.80	0.55	1416	0.581
ISAA, γ_{40}	-21.36	15.16	-1.40	1416	0.159
ISOTHER, γ_{50}	17.95	9.77	1.83	1416	0.066
ISEL, γ_{60}	-4.79	2.60	-1.83	1416	0.066
ISSWD, γ_{70}	-4.53	4.66	-0.97	1416	0.331
ISNSLP, γ_{80}	-1.02	2.98	-0.34	1416	0.732
DIBELSBEG, γ_{90}	1.38	0.04	31.46	1416	<0.001
INOGINST, γ_{100}	4.76	7.31	0.65	15	0.524

Table 21

Variance Components for the HLM Grade School Model

Random Effect	Standard Deviation	Variance Component	*d.f.*	χ^2	p-value
INTRCPT1, u_0	11.69	136.74	15	61.04	<0.001
INOGINST slope, u_{10}	23.07	532.32	15	51.70	<0.001
level-1, r	56.39	3180.48			

Table 22

Fixed Effects for the HLM First Grade Teacher Model

Fixed Effect	Coefficient	Standard error	*t*-ratio	Approx. *d.f.*	*p*-value
INTRCPT, γ_{00}	16.81	6.14	2.73	47	0.009
ISFEMALE, γ_{10}	2.70	3.02	0.89	1352	0.372
ISASIAN, γ_{20}	-25.52	10.20	-2.50	1352	0.012
ISWHITE, γ_{30}	4.45	4.52	0.98	1352	0.325
ISAA, γ_{40}	-22.95	11.33	-2.02	1352	0.043
ISOTHER, γ_{50}	20.40	11.46	1.78	1352	0.075
ISEL, γ_{60}	-4.46	3.18	-1.40	1352	0.161
ISSWD, γ_{70}	-6.06	6.22	-0.97	1352	0.330
ISNSLP, γ_{80}	-0.40	2.84	-0.14	1352	0.887
DIBELSBEG, γ_{90}	1.41	0.03	37.54	1352	<0.001
INOGINST, γ_{100}	7.38	5.30	1.39	47	0.171

Table 23

Variance Components for the HLM Grade Teacher Model

Random Effect	Standard Deviation	Variance Component	*d.f.*	χ^2	*p*-value
INTRCPT1, u_0	18.29	334.55	46	173.79	<0.001
INOGINST slope, u_{10}	22.85	522.39	46	86.51	<0.001
level-1, r	54.50	2971.22			

Table 24

Covariates before and after matching Kindergarten

	Before Matching		After Matching	
	Control	*Treatment*	*Control*	*Treatment*
% Female	52.8	45.2	40.1	45.2
% Asian	0.1	0.5	0.5	0.5
% White	16.1	11.4	8.0	11.4
% African American	2.5	1.4	0.4	1.4
% Other Ethnicity	2.0	1.7	0.4	1.7
% Hispanic	79.3	85.0	90.7	85.0
% English Learner	33.0	46.9	61.9	46.9
% Special Education	4.7	8.7	8.2	8.7
% SED	42.1	48.0	42.6	48.0
Avg. BOY Composite	32.9	1.8	1.6	1.8

Note: The sample used for the propensity score analysis included students from schools that did not implement the OG Institute. The characteristics of these schools can be seen in Table 1.

Table 25

Covariates before and after matching First Grade

	Before Matching		After Matching	
	Control	*Treatment*	*Control*	*Treatment*
% Female	49.0	41.0	41.1	41.0
% Asian	1.6	0.3	0.1	0.3
% White	25.6	9.1	5.5	9.1
% African American	1.8	1.7	1.5	1.7
% Other Ethnicity	2.1	1.0	1.5	1.0
% Hispanic	31.1	87.9	91.4	87.9
% English Learner	25.4	44.4	45.0	44.4
% Special Education	7.6	15.9	12.5	15.9
% SED	33.8	48.1	51.3	48.1
Avg. BOY Composite	119.8	54.7	56.1	54.7

Note: The sample used for the propensity score analysis included students from schools that did not implement the OG Institute. The characteristics of these schools can be seen in Table 1.

Table 26

Strata and differences for Kindergarten

Strata	N		Propensity Score Range	EOY Composite		Difference
	Control	*Treatment*		*Control*	*Treatment*	
1	488	64	[0.035, 0.106)	113.4	118.3	6.79
2	268	53	(0.106, 0.176]	104.2	107.5	0.09
3	229	61	(0.176, 0.247]	98.4	96.4	4.63
4	298	96	(0.247, 0.317]	91.9	99.1	0.80
5	88	48	(0.317, .387]	87.1	88.9	2.62
Total	**1407**	**358**		**Treatment Effect**		**3.8**

Table 27

Strata and differences for First Grade

Strata	N		Propensity Score Range	EOY Composite		Difference
	Control	*Treatment*		*Control*	*Treatment*	
1	1472	143	[0.024, 0.168)	125.4	128.5	3.1
2	180	84	(0.168, 0.312]	66.7	72.5	5.8
3	93	32	(0.312, 0.456]	52.0	55.8	3.8
4	45	26	(0.456, 0.600]	31.8	54.1	22.2
5	22	10	(0.600, .744]	12.3	43.9	31.6
Total	**1812**	**295**		**Treatment Effect**		**4.6**

Table 28

Short Term Outcomes of OG Institute Students Kindergarten

	Well-Below Benchmark		Below Benchmark		At Benchmark		Total
	N	**%**	**N**	**%**	**N**	**%**	
Cohort 1	32	32%	36	36%	33	33%	101
Cohort 2	19	17%	19	17%	73	66%	111
Cohort 3	104	71%	33	23%	9	6%	146
Total	**155**	**43%**	**88**	**25%**	**115**	**32%**	**358**

Note: Students that participated in the OG Institute that had DIBELS EOY scores in the 2014-15 school year

Table 29

Short Term Outcomes of OG Institute Students First Grade

	Well-Below Benchmark		Below Benchmark		At Benchmark		Total
	N	**%**	**N**	**%**	**N**	**%**	
Cohort 1	62	70%	8	9%	19	21%	89
Cohort 2	52	46%	15	13%	45	40%	112
Cohort 3	52	55%	13	14%	29	31%	94
Total	**166**	**56%**	**36**	**12%**	**93**	**32%**	**295**

Note: Students that participated in the OG Institute that had DIBELS EOY scores in the 2014-15 school year

FIGURES

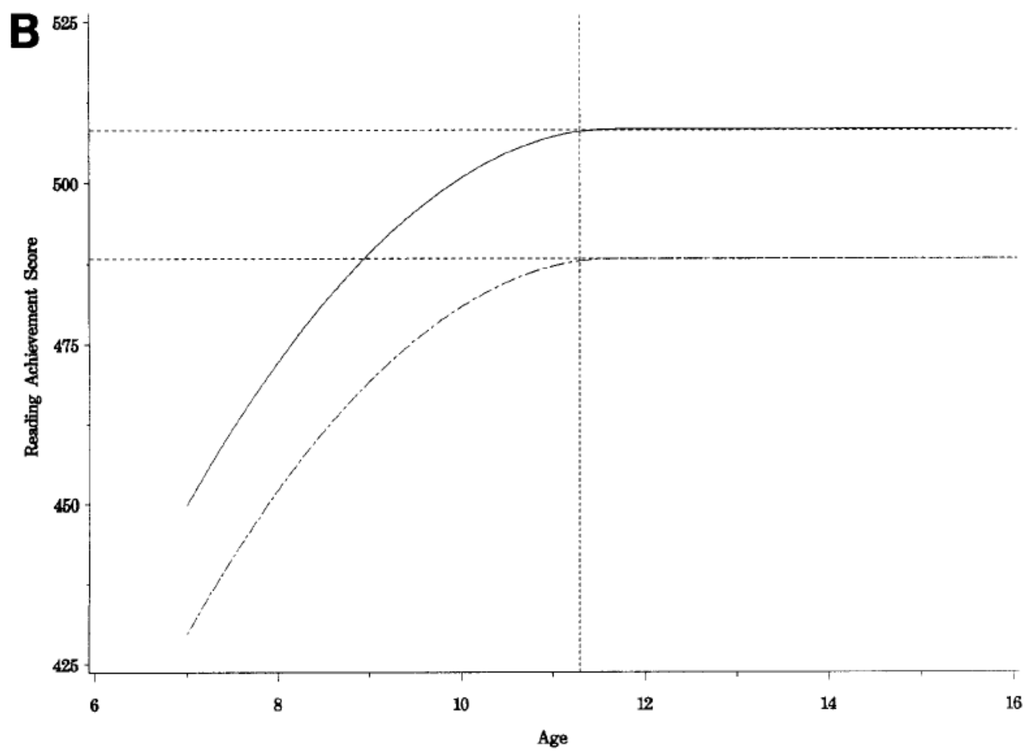

Figure 1: The quadratic to plateau model theorized by Francis & Shaywitz (1983) showing the deficit model of reading growth.

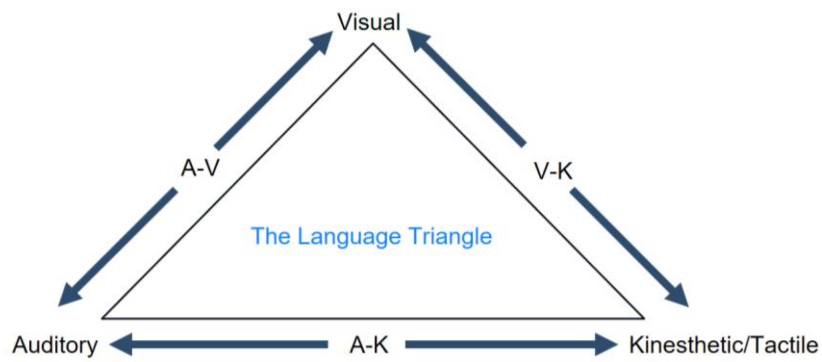

Figure 2. The Language Triangle as seen in the Orton-Gillingham Manual

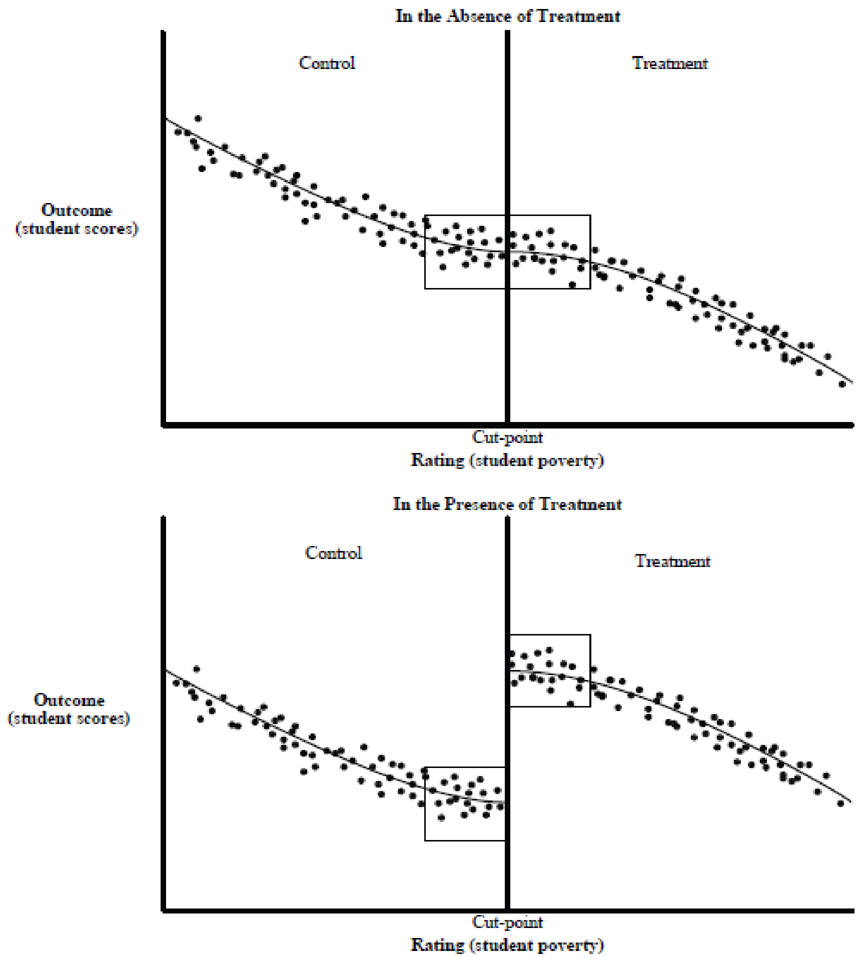

Figure 3: A graphical representation of a regression discontinuity from Cook & Shadish (2009). In this example, students above a certain poverty rating are put in an intervention. The top graph shows that a smooth relationship exists between the rating variable and the outcome when no treatment is in place. In the presence of the intervention, the jump in scores at the cut-point is a direct measurement of the treatment effect.

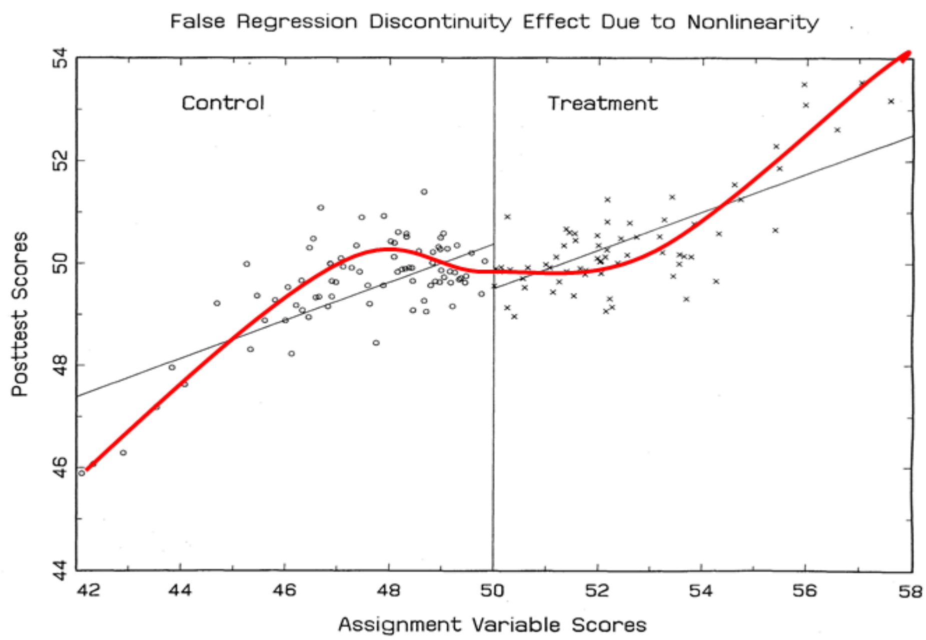

Figure 4. An illustration of the importance of functional form in an RD design. If the form is misspecified, as in this graph, a treatment effect can be falsely attributed to the intervention. From Cook & Shadish (2009)

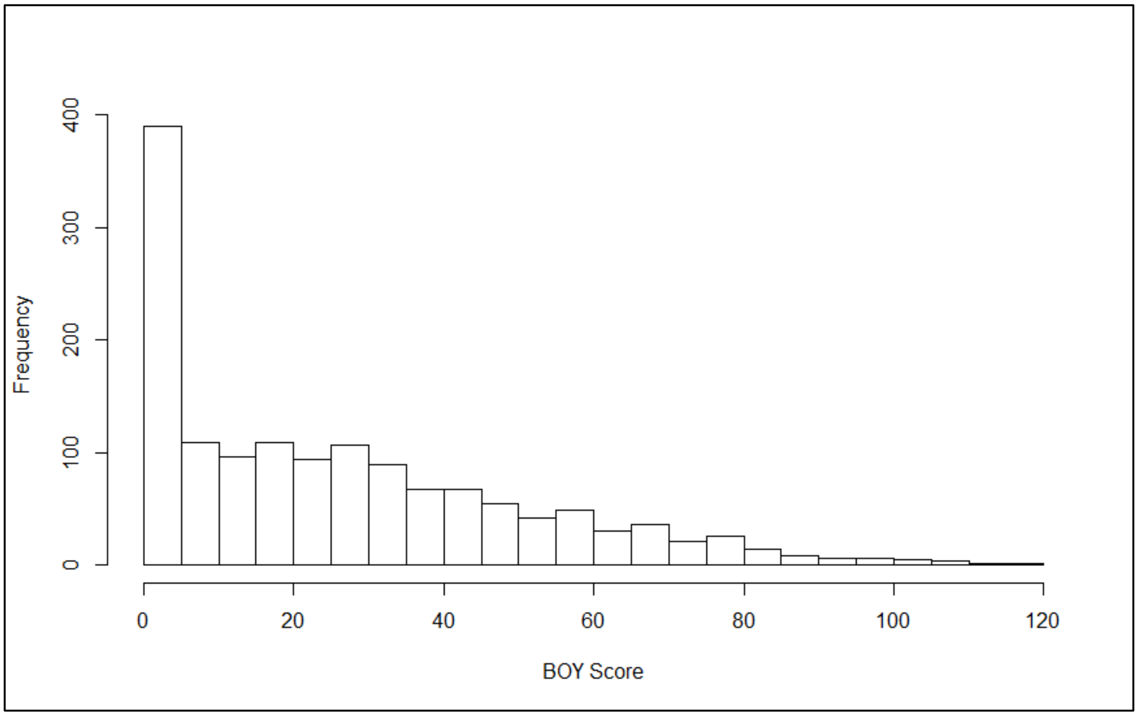

Figure 5. Histogram of the Kindergarten Beginning of Year (BOY) Composite Scores.

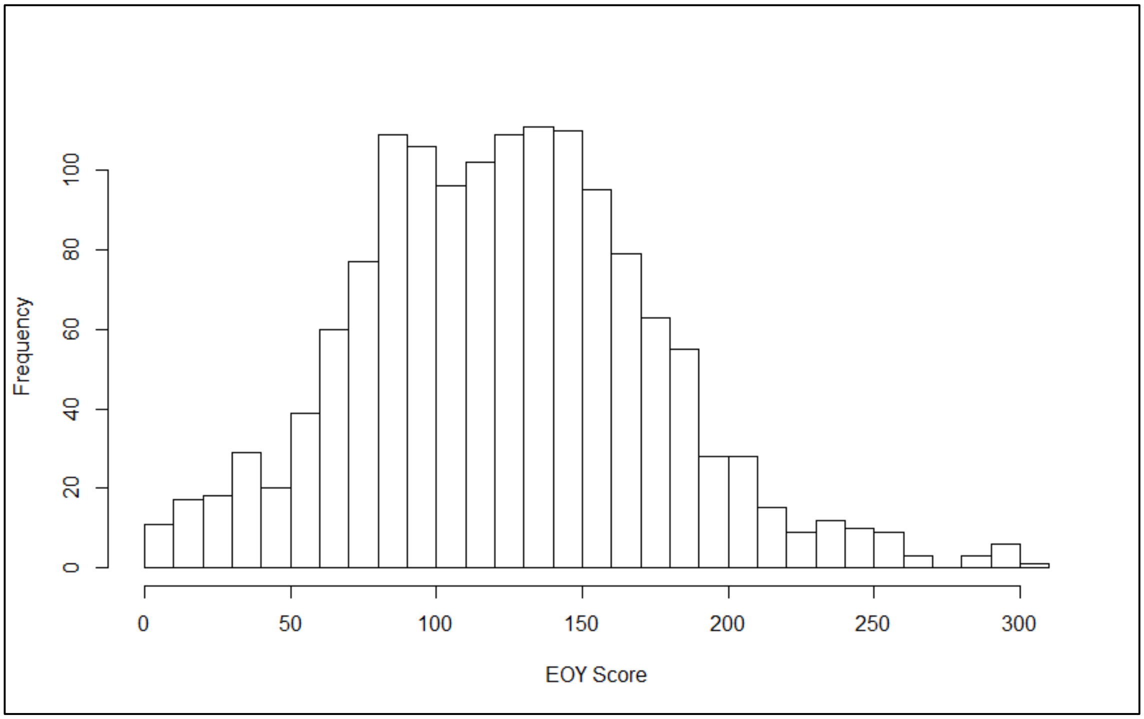

Figure 6. Histogram of the Kindergarten End of Year (EOY) Composite Scores.

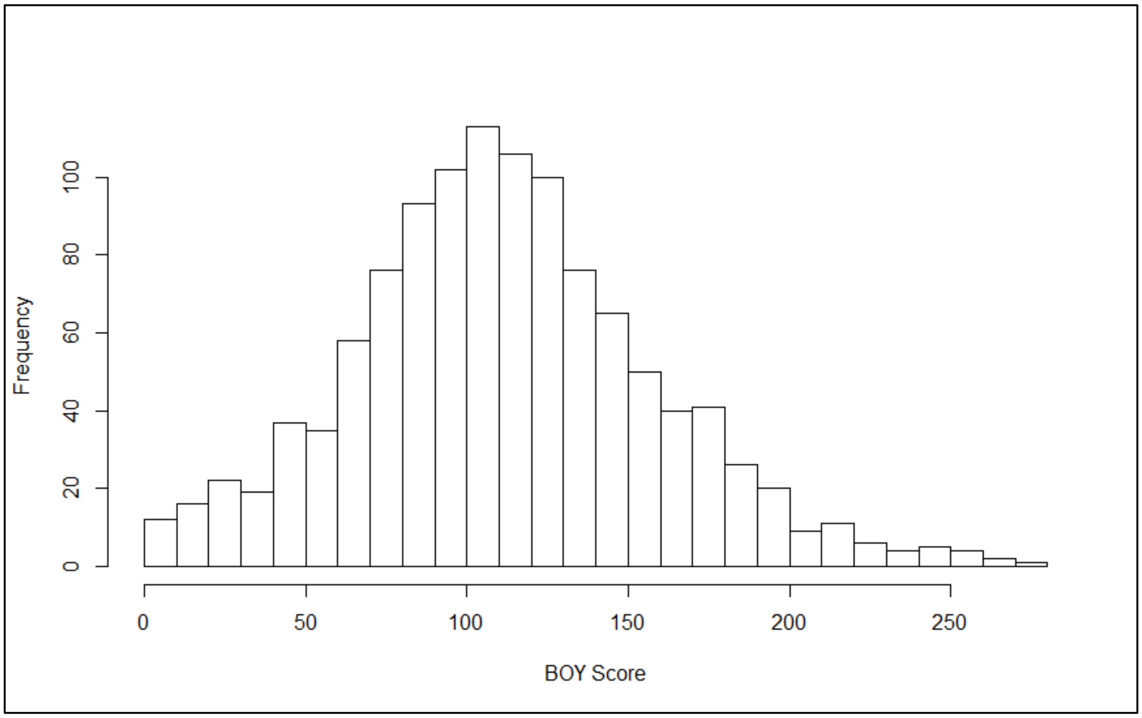

Figure 7. Histogram of the Beginning of Year (BOY) Composite Scores for First Grade.

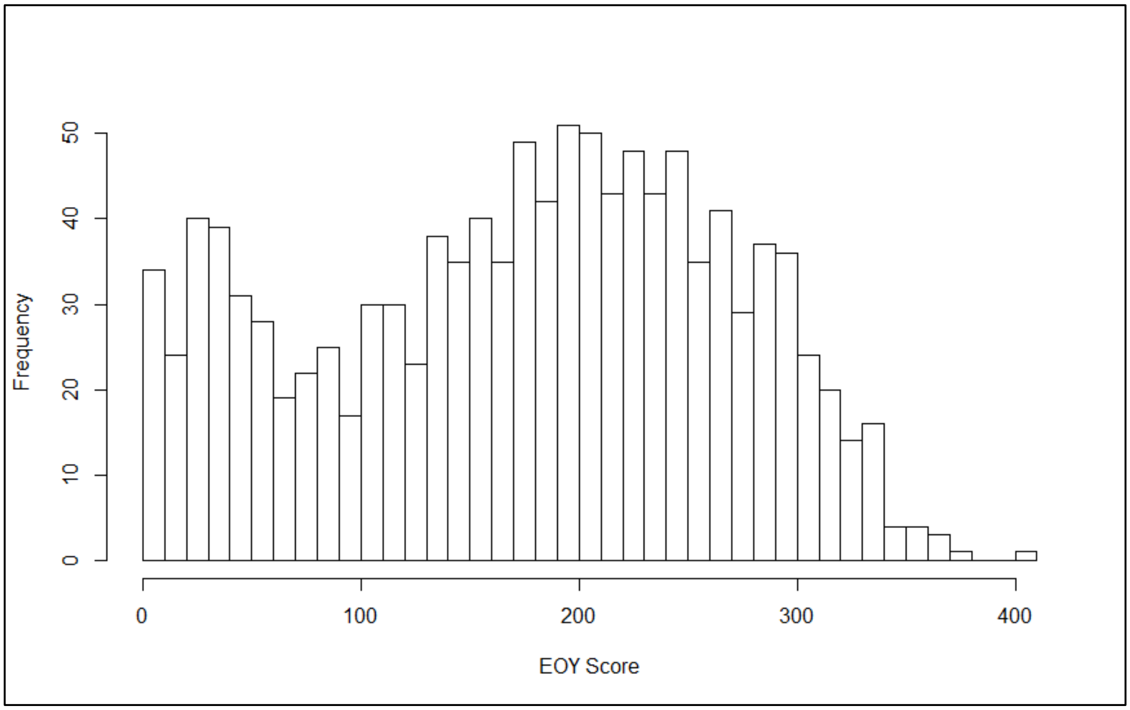

Figure 8. Histogram of the End of Year (EOY) Composite Scores for First Grade.

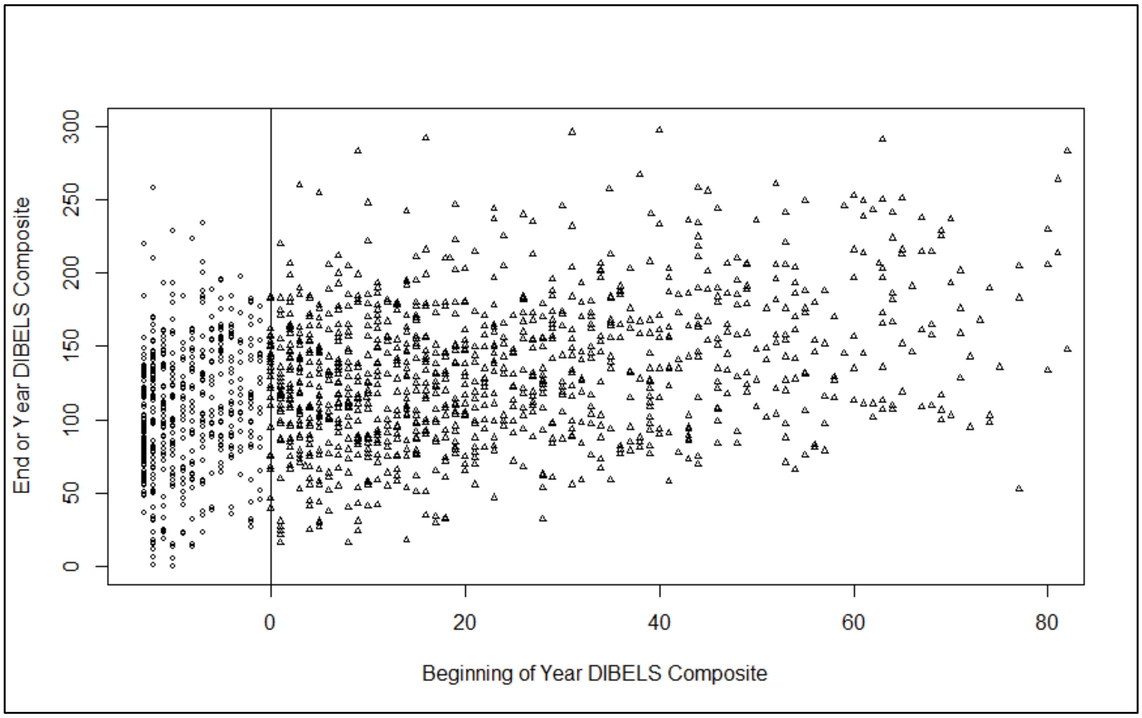

Figure 9. Scatterplot of Kindergarten BOY Composite and EOY Composite Scores at OG Institute Schools.

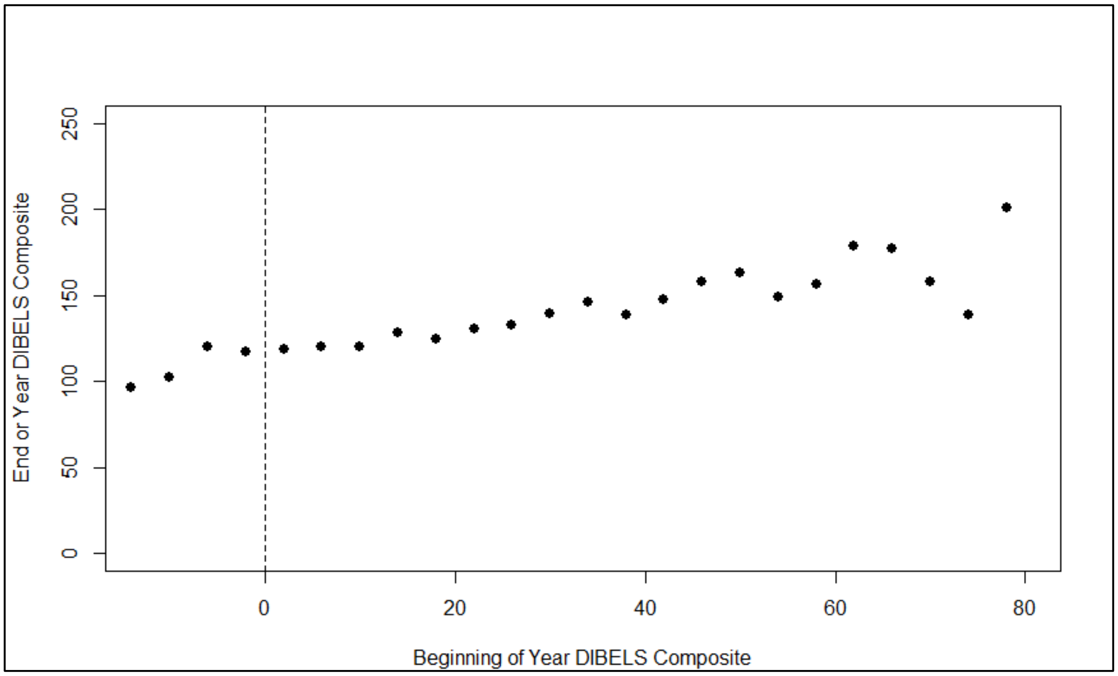

Figure 10. Binned Scatterplot of Kindergarten BOY and EOY scores at OG Institute Schools.

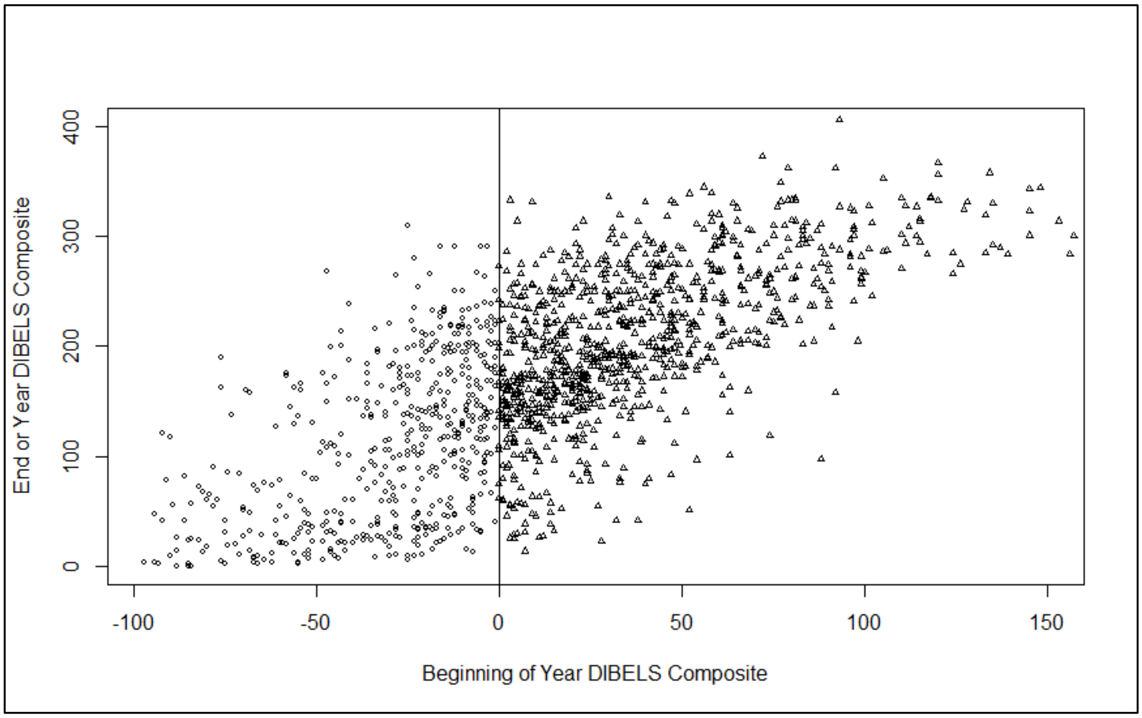

Figure 11. Scatterplot of First Grade BOY Composite and EOY composite scores at OG Institute Schools.

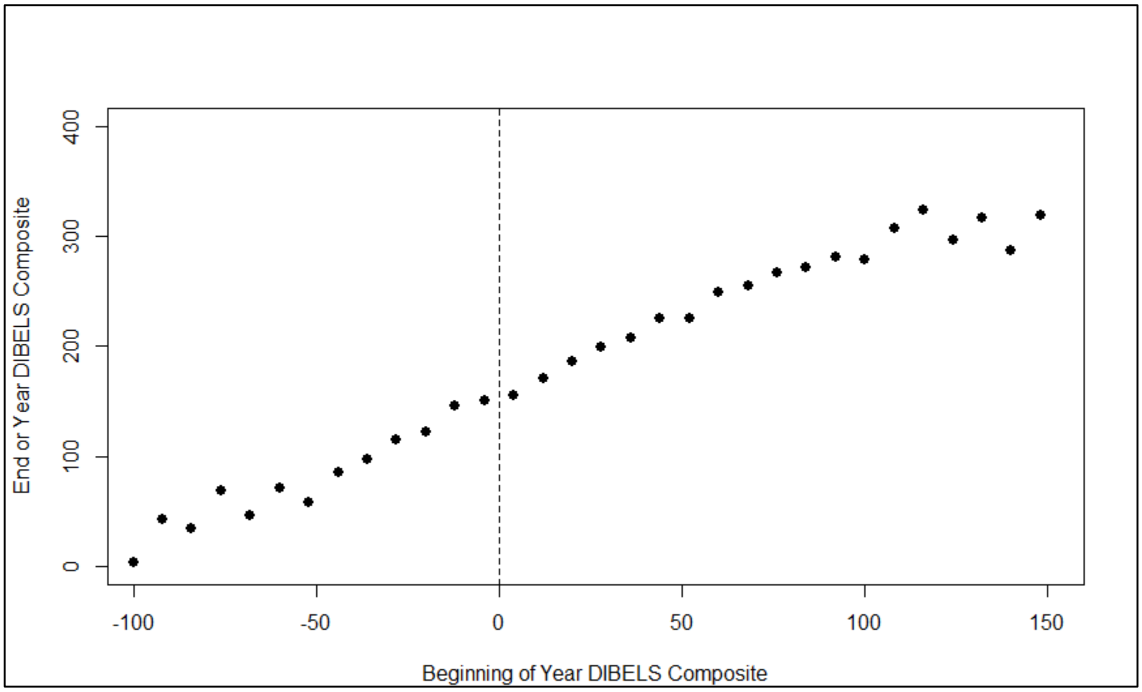

Figure 12. Binned Scatterplot of First Grade BOY and EOY scores at OG Institute Schools.

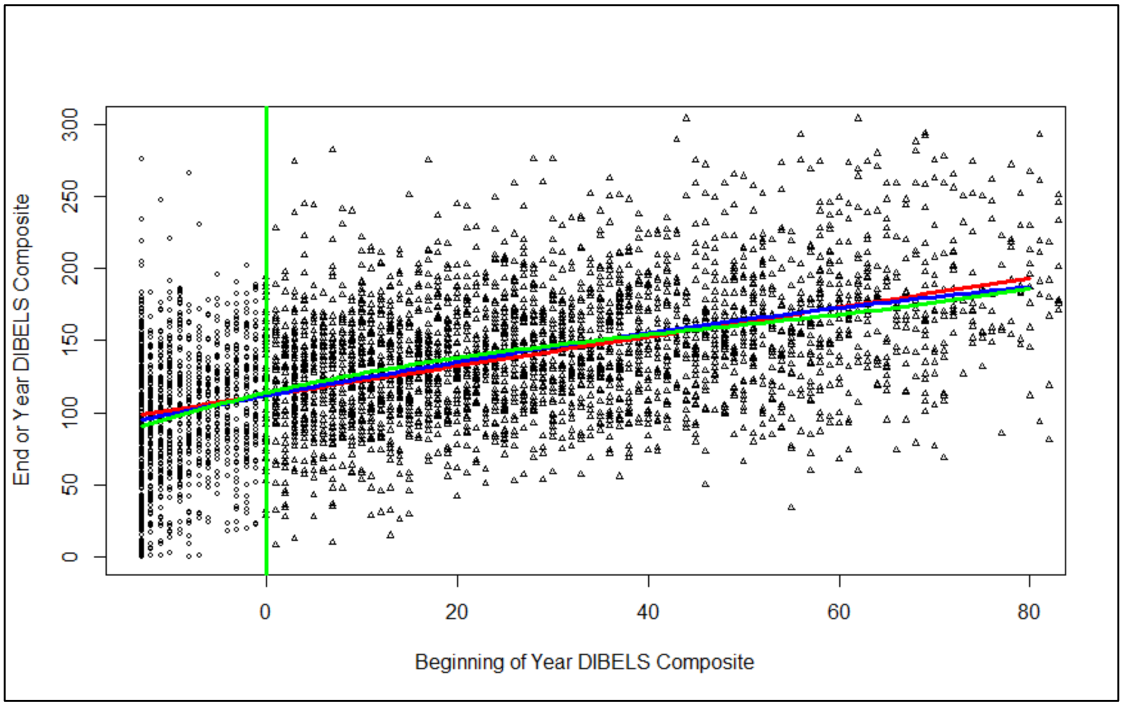

Figure 13. Scatterplot of Kindergarten BOY Composite and EOY Composite Scores at schools that did not implement the OG Institute. This sample was used to estimate the parametric model that best fits the relationship between BOY and EOY composite scores.

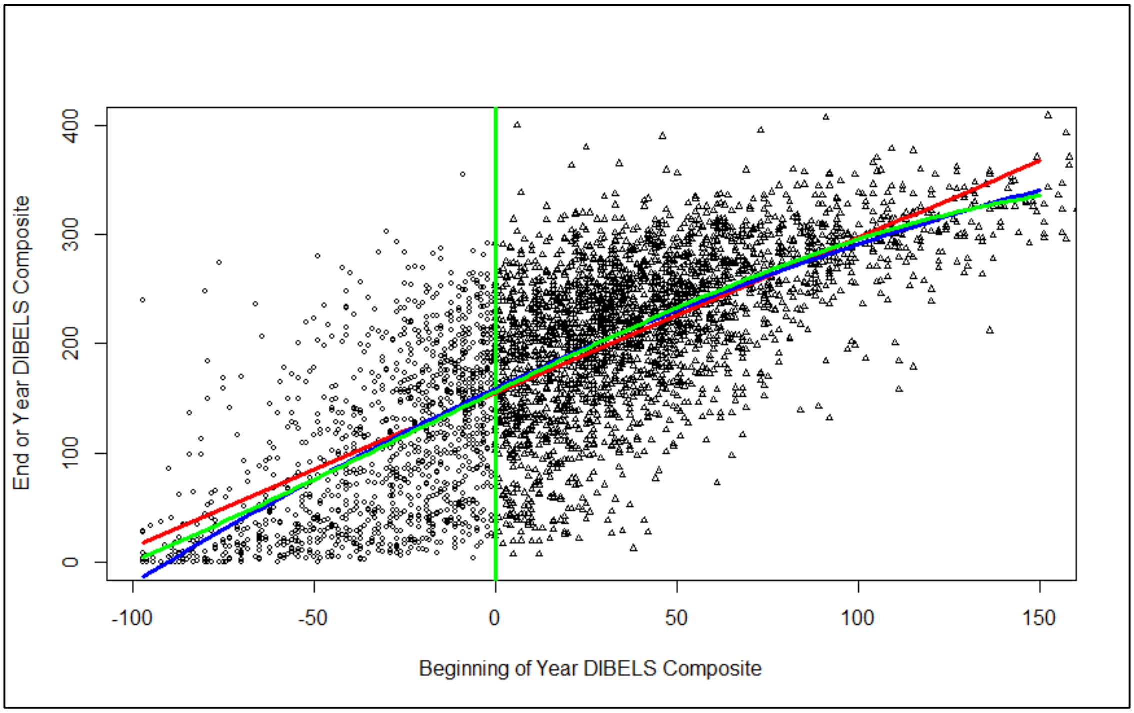

Figure 14. Scatterplot of First Grade BOY Composite and EOY Composite Scores at schools that did not implement the OG Institute. This sample was used to estimate the parametric model that best fits the relationship between BOY and EOY composite scores.

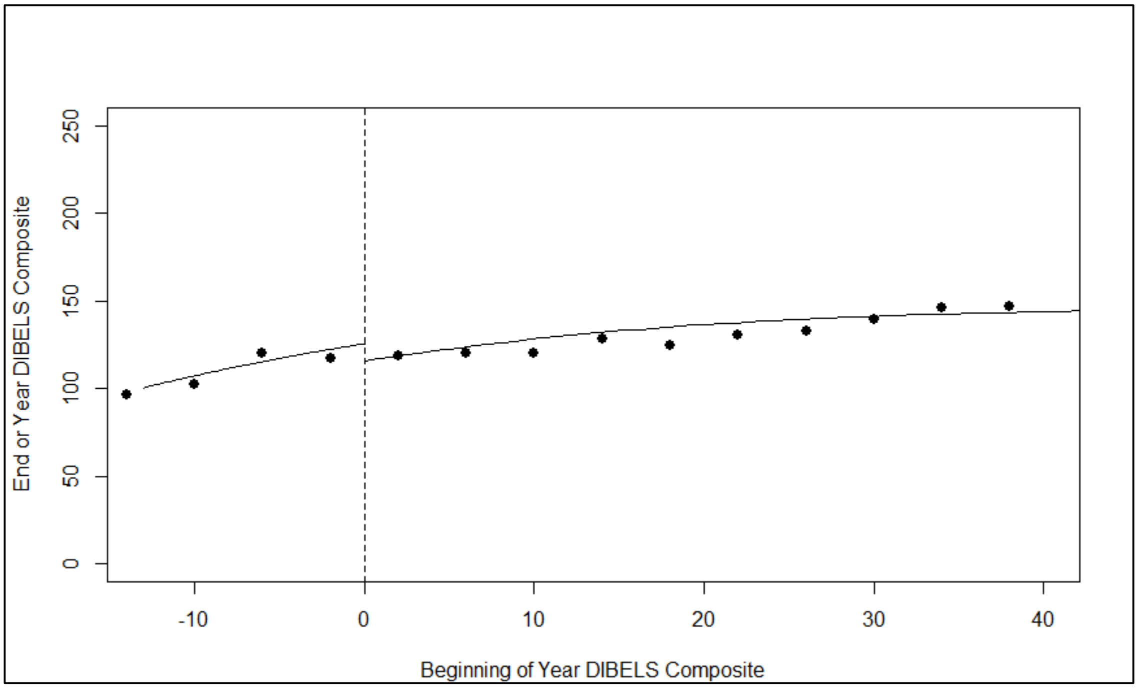

Figure 15. Graph of the kindergarten RD parametric analysis. The cubic function is fitted to the binned data, showing a non-significant treatment effect of 8.79 composite score points.

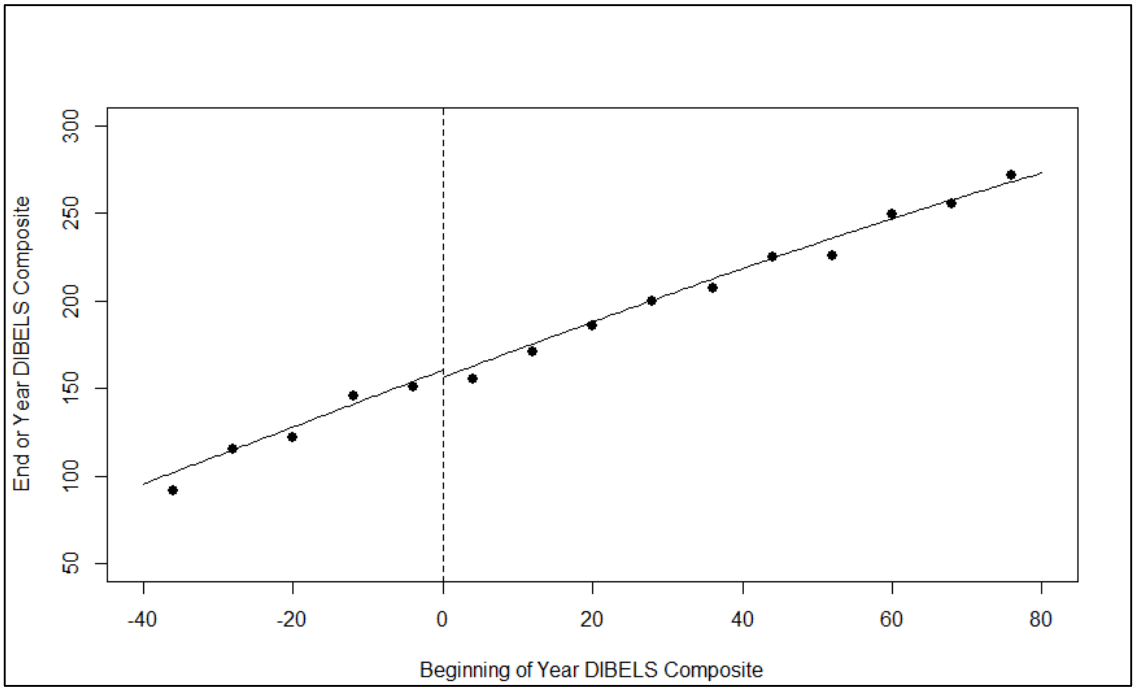

Figure 16. Graph of the first grade RD parametric analysis. The cubic function is fitted to the binned data, showing a treatment effect (non-significant) of 3.8 composite score points.

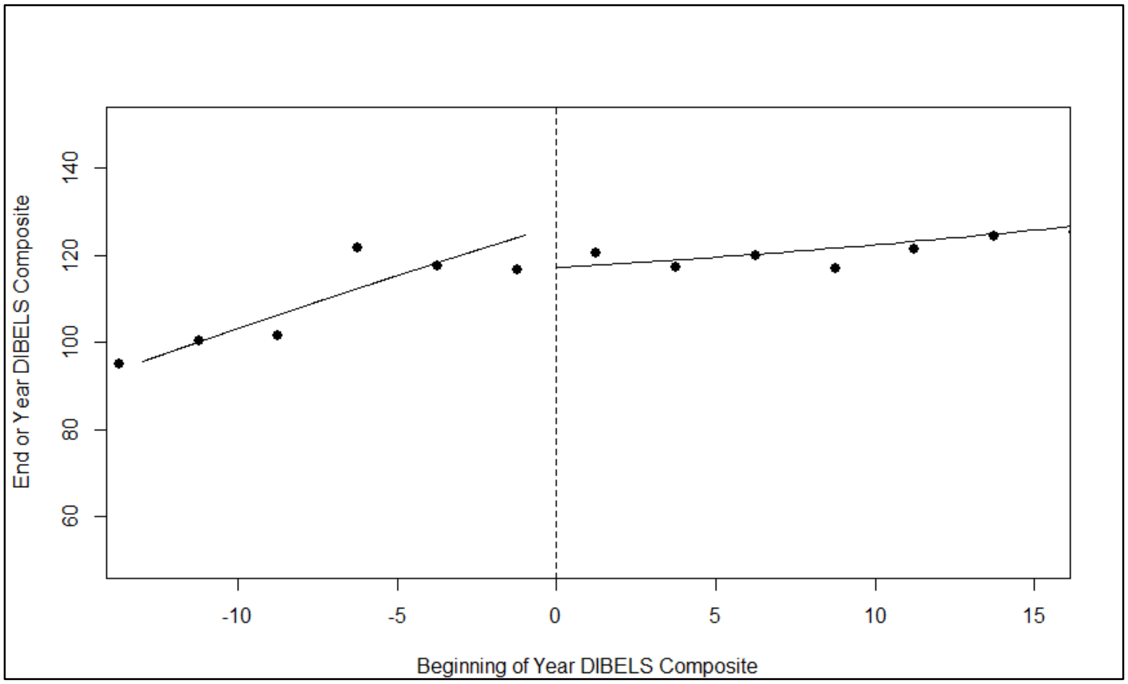

Figure 17. Graph of the kindergarten RD non-parametric analysis. The local linear regression is fitted to the binned data, showing a treatment effect (non-significant) of 6.7 composite score points.

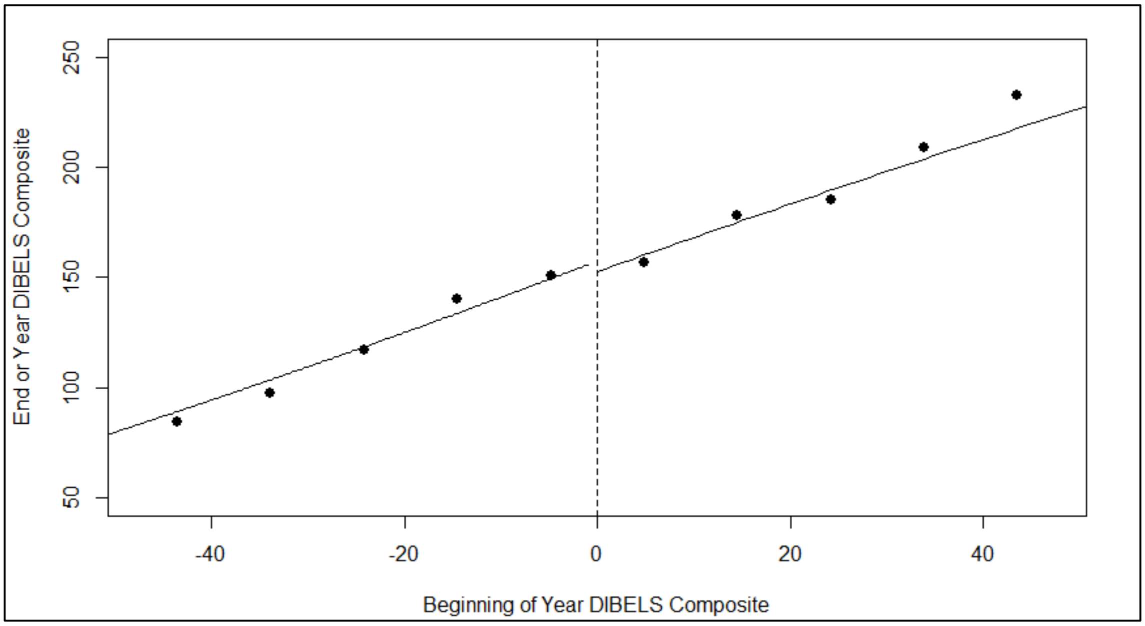

Figure 18. Graph of the first grade RD non-parametric analysis. The local linear regression is fitted to the binned data, showing a treatment effect (non-significant) of 2.3 composite score points.

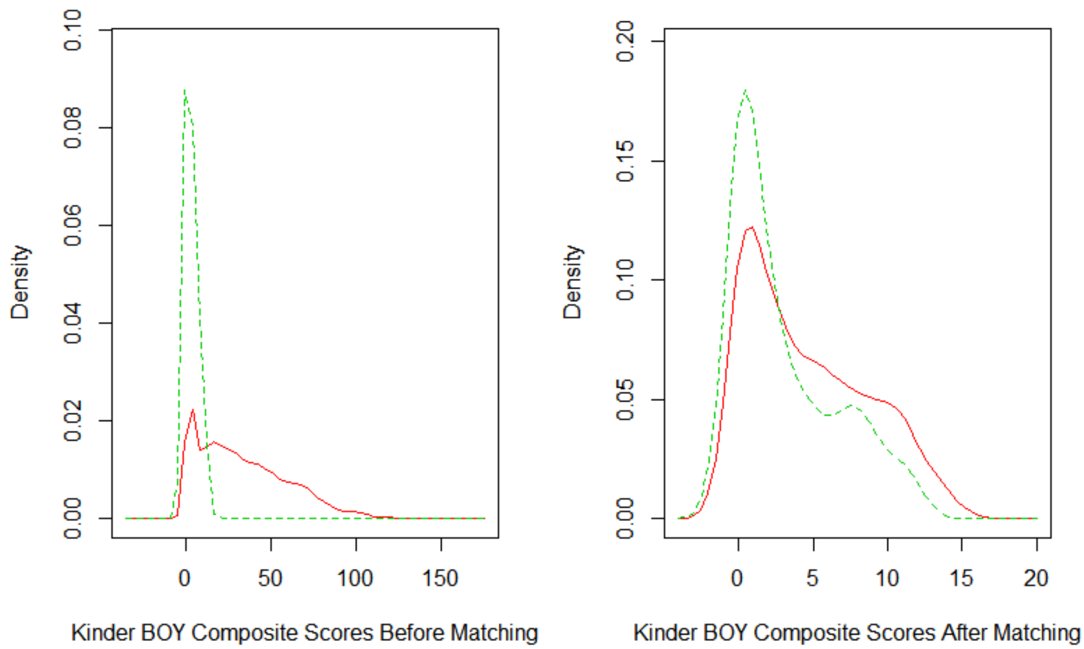

Figure 19. Kernel Density Plots of Kindergarten BOY Composite Scores before and after matching.

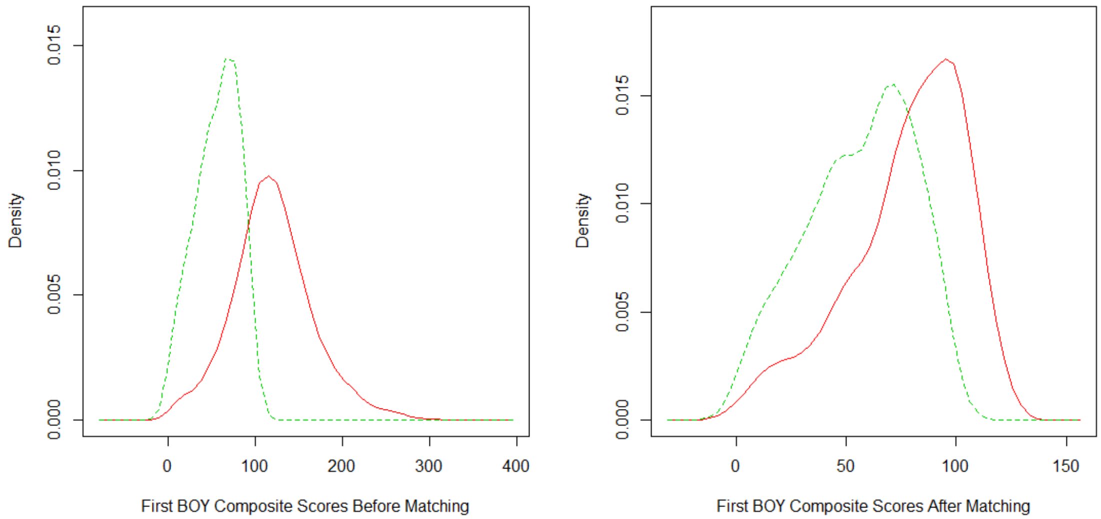

Figure 20. Kernel Density Plots of First Grade BOY Composite Scores before and after matching.

References

Acalin, T. A. (1995). A comparison of Reading Recovery to Project Read (Master's thesis, California State University, Fullerton, 1995). Masters Abstracts International, 33(06), 54-1660.

Adams, M. J. (1990). Beginning to read: Learning and thinking about print. *Beginning to read: Learning and thinking about print*.

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J. L., & Hsiung, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation efficacy of a university-level remedial writing program.Evaluation Review, 22(2), 207-244.

Alexander, A. W., & Slinger-Constant, A. M. (2004). Current status of treatments for dyslexia: Critical review. *Journal of Child Neurology*, *19*(10), 744-758.

Anthony, J. L., & Francis, D. J. (2005). Development of phonological awareness. *Current Directions in Psychological Science*, *14*(5), 255-259.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, *46*(3), 399-424.

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

Berk, R., Barnes, G., Ahlman, L., & Kurtz, E. (2010). When second best is good enough: A comparison between a true experiment and a regression discontinuity quasi-experiment. Journal of Experimental Criminology, 6(2), 191-208.

Birsh, J. R. (2005). Multisensory structured language education. *Multisensory teaching of basic language skills*,

Bitter, G., & White, M. A. (2011). Final summary report: Evaluation study of the Writing Road to Reading. Scottsdale: Arizona State University.

Blachman, B. A., Ball, E. W., Black, R. S., & Tangel, D. M. (1994). Kindergarten teachers develop phoneme awareness in low-income, inner-city classrooms. *Reading and Writing*, *6*(1), 1-18.

Blachman, B. A., Tangel, D. M., Ball, E. W., Black, R., & McGraw, C. K. (1999). Developing phonological awareness and word recognition skills: A two-year intervention with low-income, inner-city children. *Reading and Writing*, *11*(3), 239-273.

Black, D., Galdo, J., & Smith, J. A. (2007). Evaluating the bias of the regression discontinuity design using experimental data. University of Chicago, Working paper.

Bloom, H. S. (2012). Modern regression discontinuity analysis. Journal of Research on Educational Effectiveness, 5(1), 43-82.

Bowey, J. A. (2006). Need for systematic synthetic phonics teaching within the early reading curriculum. *Australian Psychologist*, *41*(2), 79-84.

Bowyer-Crane, C., Snowling, M. J., Duff, F. J., Fieldsend, E., Carroll, J. M., Miles, J., ... & Hulme, C. (2008). Improving early language and literacy skills: Differential effects of an oral language versus a phonology with reading intervention. *Journal of Child Psychology and Psychiatry*, *49*(4), 422-432.

Bradley, L., & Bryant, P. E. (1983). Categorizing sounds and learning to read: A causal connection. *Nature*, 419-421.

Brady, S. A. (2011). Efficacy of phonics teaching for reading outcomes. *Explaining individual differences in reading: Theory and evidence*, 69.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. (2010). HLM 7 for Windows [Computer software]. Chicago, IL: Scientific Software International.

Cameron, A. C., & Miller, D. L. (2015). A practitioner’s guide to cluster-robust inference. Journal of Human Resources, 50(2), 317-372.

Cohen, J. (1992). A power primer. Psychological bulletin, 112(1), 155.

Duncan, G. J., Brooks-Gunn, J., & Klebanov, P. K. (1994). Economic deprivation and early childhood development. Child development, 65(2), 296-318.

Baker, S. K., Smolkowski, K., Chaparro, E. A., Smith, J. L., & Fien, H. (2014). Using Regression Discontinuity to Test the Impact of a Tier 2 Reading Intervention in First Grade. Journal of Research on Educational Effectiveness, (just-accepted), 00-00.

Becker, S. O., & Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. The stata journal, 2(4), 358-377.

Bryant, D. P., Bryant, B. R., Gersten, R., Scammacca, N., & Chavez, M. M. (2008). Mathematics Intervention for First-and Second-Grade Students With Mathematics

Difficulties The Effects of Tier 2 Intervention Delivered as Booster Lessons. Remedial and Special Education, 29(1), 20-32.

Buddelmeyer, H., & Skoufias, E. (2004). An evaluation of the performance of regression discontinuity design on PROGRESA (Vol. 827). World Bank Publications.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. Sociological methods & research, 33(2), 261-304.

Bursuck, W. D., & Damer, M. (2007). *Reading instruction for students who are at risk or have disabilities*. Pearson/Allyn and Bacon.

Bus, A. G., & van IJzendoorn, M. H. (1999). Phonological awareness and early reading: A meta-analysis of experimental training studies. *Journal of Educational Psychology*, *91*(3), 403.

Bussjaeger, J. J. (1993). The effectiveness of Project Read on the reading achievement of students with learning disabilities. (Master's thesis, California State University, Fullerton, 1993). Masters Abstracts International, 31 (04), 54-1480.

Byrne, B., & Fielding-Barnsley, R. (1991). Evaluation of a program to teach phonemic awareness to young children. *Journal of Educational Psychology*, *83*(4), 451.

Cain, K., & Oakhill, J. (2011). Matthew effects in young readers reading comprehension and reading experience aid vocabulary development. *Journal of Learning Disabilities*, *44*(5), 431-443.

Catts, H. W., Bridges, M. S., Little, T. D., & Tomblin, J. B. (2008). Reading achievement growth in children with language impairments. *Journal of Speech, Language, and Hearing Research*, *51*(6), 1569-1579.

Catts, H. W., Petscher, Y., Schatschneider, C., Bridges, M. S., & Mendoza, K. (2008). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. Journal of Learning Disabilities.

Christensen, C. A., & Bowey, J. A. (2005). The efficacy of orthographic rime, grapheme–phoneme correspondence, and implicit phonics approaches to teaching decoding skills. *Scientific Studies of Reading*, *9*(4), 327-349.

Clearinghouse, W. W. Institute of Education Sciences.(2010). *Orton-Gillingham based strategies unbranded*.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics, 295-313.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. (1966). Equality of educational opportunity.Washington, dc, 1066-5684.

Cook, T. D. (2008). “Waiting for life to arrive”: a history of the regression-discontinuity design in psychology, statistics and economics. Journal of Econometrics, 142(2), 636-654.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. Journal of policy analysis and management, 27(4), 724-750.

Cook, T. D., & Steiner, P. M. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of pretest measures of outcome, of unreliable measurement, and of mode of data analysis. *Psychological methods*, *15*(1), 56.

Connor, C. M., Morrison, F. J., & Katch, L. E. (2004). Beyond the reading wars: Exploring the effect of child-instruction interactions on growth in early reading. *Scientific Studies of Reading*, *8*(4), 305-336.

Cunningham, A. E. (1990). Explicit versus implicit instruction in phonemic awareness. *Journal of experimental child psychology*, *50*(3), 429-444.

d’Agostino, R. B. (1998). Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*, *17*(19), 2265-2281.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, *84*(1), 151-161.

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.

DesJardins, S., & McCall, B. (2008). The impact of the Gates Millennium Scholars Program on the retention, college finance-and work-related choices, and future educational aspirations of low-income minority students. Unpublished Manuscript.

Dewey, E. N., Good, R. H., & Ph III, D. (2014). DIBELS Next® National Norms 2012--2013 (No. 17). Technical Report.

Donovan, S., & Cross, C. (2002). Minority students in gifted and special education. *National Research Council, Washington DC*.

Ehri, L. C., Dreyer, L. G., Flugman, B., & Gross, A. (2007). Reading rescue: An effective tutoring intervention model for language-minority students who are struggling readers in first grade. American Educational Research Journal, 44(2), 414–48.

Felton, R. H. (1992). Early identification of children at risk for reading disabilities. *Topics in Early Childhood Special Education*, *12*(2), 212-229.

Fernald, G. M., & Keller, H. (1921). The effect of kinaesthetic factors in the development of word recognition in the case of non-readers. *The Journal of Educational Research*, *4*(5), 355-377.

Fien, H., Smith, J. L. M., Smolkowski, K., Baker, S. K., Nelson, N. J., & Chaparro, E. (2014). An Examination of the Efficacy of a Multitiered Intervention on Early Reading Outcomes for First Grade Students at Risk for Reading Difficulties. Journal of learning disabilities, 0022219414521664.

Fletcher, J. M., Coulter, W. A., Reschly, D. J., & Vaughn, S. (2004). Alternative approaches to the definition and identification of learning disabilities: Some questions and answers. *Annals of Dyslexia*, *54*(2), 304-331.

Fletcher, J. M., Lyon, G. R., Fuchs, L. S., & Barnes, M. A. (2006). *Learning disabilities: From identification to intervention*. Guilford Press.

Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology*, *90*(1), 37.

Fox, B., & Routh, D. K. (1984). Phonemic analysis and synthesis as word attack skills: Revisited. *Journal of Educational Psychology*, *76*(6), 1059.

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational psychology*, *88*(1), 3.

Fuchs, L. S., & Fuchs, D. (1998). Treatment validity: A unifying concept for reconceptualizing the identification of learning disabilities. Learning Disabilities Research & Practice.

Fuchs, D., Fuchs, L. S., & Compton, D. L. (2004). Identifying reading disabilities by responsiveness-to-instruction: Specifying measures and criteria.Learning Disability Quarterly, 27(4), 216-227.

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Walters, K., Song, M., ... & Doolittle, F. (2010). Middle School Mathematics Professional Development Impact Study: Findings After the First Year of Implementation. NCEE 2010-4009. National Center for Education Evaluation and Regional Assistance.

Gersten, R., Compton, D., Connor, C. M., Dimino, J., Santoro, L., Linan-Thompson, S., & Tilly, W. D. (2009) Assisting Students Struggling with Reading: Response to Intervention (RtI) and Multi-Tier Intervention in the Primary Grades. Retreived from http://ies. ed. gov/ncee/wwc.*PracticeGuide.aspx*.

Gillingham, A., & Stillman, B. W. (1997). *The gillingham manual: Remedial training for students with specific disability in reading, spelling, and penmanship*. Educators Pub. Service.

Gleason, P. M., Resch, A. M., & Berk, J. A. (2012). Replicating experimental impact estimates using a regression discontinuity approach (No. c7ecf8b35c8d402491d09b2d0823e99f). Mathematica Policy Research.

Good, R. H., Kaminski, R. A., Cummings, K., Dufour-Martel, C., Peterson, K., Powell-Smith, K., & Wallin, J. (2011). DIBELS next assessment manual. Eugene, OR: Dynamic Measurement Group.

Good, R. H., Kaminski, R. A., Dewey, E. N., & Wallin, J. Powell--Smith, KA, Latimer, R., J.,(2013). *DIBELS Next Technical Manual*.

Good III, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*, *5*(3), 257-288.

Gough, P. B., & Hillinger, M. L. (1980). Learning to read: An unnatural act. *Annals of Dyslexia*, *30*(1), 179-196.

Giess, S., Rivers, K. O., Kennedy, K., & Lombardino, L. J. (2012). Effects of Multisensory Phonics-Based Training on the Word Recognition and Spelling Skills of Adolescents with Reading Disabilities. *International Journal of Special Education, 27*(1), 60-73.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. Annual review of psychology, 60, 549-576.

Griffin, P., Burns, M. S., & Snow, C. E. (Eds.). (1998). *Preventing reading difficulties in young children*. National Academies Press.

Gunn, B., Biglan, A., Smolkowski, K., & Ary, D. (2000). The efficacy of supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school. The Journal of Special Education, 34(2), 90–103.

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica,69(1), 201-209.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.

Hattie, John. *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge, 2013.

Henry, M. K. (1998). Structured, sequential, multisensory teaching: The Orton legacy. *Annals of Dyslexia*, *48*(1), 1-26.

Hishinuma, E. S., & Nishimura, S. T. (2000). Parent attitudes on the importance and success of integrated self-contained services for students who are gifted, learning disabled, and gifted/learning disabled. *Roeper Review*, *22*(4), 241-250.

Hollands, F. M., Pan, Y., Shand, R., Cheng, H., Levin, H. M., Belfield, C. R., ... & Hanisch-Cerda, B. (2013). improving early literacy: cost-effectiveness analysis of effective reading programs. New York: Center for Benefit-Cost Studies of Education, Teachers College, Columbia University.

Hong, G., & Raudenbush, S. W. (2008). Causal inference for time-varying instructional treatments. Journal of Educational and Behavioral Statistics,33(3), 333-362.

Hulme, C., Bowyer-Crane, C., Carroll, J. M., Duff, F. J., & Snowling, M. J. (2012). The causal role of phoneme awareness and letter-sound knowledge in learning to read combining intervention studies with mediation analyses. *Psychological Science*, *23*(6), 572-577.

Hwee, N. C. K., & Houghton, S. (2011). The effectiveness of Orton-Gillingham-based instruction with Singaporean children with specific reading disability (dyslexia). *British Journal of Special Education*, *38*(3), 143-149.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. Journal of econometrics, 142(2), 615-635.

Imbens, G., & Kalyanaraman, K. (2011). Optimal bandwidth choice for the regression discontinuity estimator. The Review of Economic Studies, rdr043.

Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. Review of economics and statistics, 86(1), 226-244.

Jacob, R. T., Zhu, P., Somers, M. A., & Bloom, H. S. (2012). A practical guide to regression discontinuity. MDRC.

Jenkins, J. R., Peyton, J. A., Sanders, E. A., & Vadasy, P. F. (2004). Effects of reading decodable texts in supplemental first-grade tutoring. Scientific Studies of Reading, 8(1), 53–85.

Johnston, R. S., & Watson, J. E. (2005). *A seven year study of the effects of synthetic phonics teaching on reading and spelling attainment*. Scottish Executive Education Department, Information, Analysis and Communication Division.

Joshi, R. M., Dahlgren, M., & Boulware-Gooden, R. (2002). Teaching reading in an inner city school through a multisensory teaching approach. *Annals of Dyslexia*, *52*(1), 229-242.

Joshi, R. M., Binks, E., Hougen, M., Dahlgren, M. E., Ocker-Dean, E., & Smith, D. L. (2009). Why elementary teachers might be inadequately prepared to teach reading. Journal of Learning Disabilities.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of educational Psychology*, *80*(4), 437.

Juel, C., & Minden-Cupp, C. (2000). Learning to read words: Linguistic units and instructional strategies. *Reading Research Quarterly*, *35*(4), 458-492.

Kutner, M., Greenberg, E., Jin, Y., Boyle, B., Hsu, Y. C., & Dunleavy, E. (2007). Literacy in Everyday Life: Results from the 2003 National Assessment of Adult Literacy. NCES 2007-490. National Center for Education Statistics.

Lee, J., Grigg, W., & Donahue, P. (2007). The nation's report card. *Reading*, *496*.

Lee, D. S., & Lemieux, T. (2009). Regression discontinuity designs in economics (No. w14723). National Bureau of Economic Research.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. Statistics in medicine, 29(3), 337-346.

Lennon, J. E., & Slesinski, C. (1999). Early intervention in reading: Results of a screening and intervention program for kindergarten students. School Psychology Review, 28(3), 353–364.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, *74*(6), 431.

Litcher, J. H., & Roberge, L. P. (1979). First grade intervention for reading achievement of high risk children. *Annals of Dyslexia*, *29*(1), 238-244.

Little, R. J., & Rubin, D. B. (2014). Statistical analysis with missing data. John Wiley & Sons.

Lonigan, C. J., & Shanahan, T. (2009). Developing Early Literacy: Report of the National Early Literacy Panel. Executive Summary. A Scientific Synthesis of Early Literacy Development and Implications for Intervention. *National Institute for Literacy*.

Lloyd, D. N. (1978). Prediction of school failure from third-grade data. *Educational and Psychological Measurement*, *38*(4), 1193-1200.

Ludwig, J., & Miller, D. L. (2005). Does Head Start improve children's life chances? Evidence from a regression discontinuity design (No. w11702). National Bureau of Economic Research.

Lundberg, I., Frost, J., & Petersen, O. P. (1988). Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading research quarterly*, 263-284.

Lyon, G. R., Fletcher, J. M., Shaywitz, S. E., Shaywitz, B. A., Torgesen, J. K., Wood, F. B., ... & Olson, R. (2001). Rethinking learning disabilities. *Rethinking special education for a new century*, 259-287.

Mathes, P. G., Denton, C., Fletcher, J., Anthony, J., Francis, D., & Schatschneider, C. (2005). The effects of theoretically different instruction and student characteristics on the skills of struggling readers. Reading Research Quarterly, 40(2), 148–182.

McBee, M. (2010). Modeling outcomes with floor or ceiling effects: An introduction to the Tobit model. Gifted Child Quarterly, 54(4), 314-320.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. Journal of Econometrics, 142(2), 698-714.

McMaster, K. L., Fuchs, D., Fuchs, L. S., & Compton, D. L. (2005). Responding to nonresponders: An experimental field trial of identification and intervention methods. Exceptional Children, 71(4), 445–463.

Montessori, M. (1912). The Montessori Method.

National Reading Panel (US), National Institute of Child Health, & Human Development (US). (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*. National Institute of Child Health and Human Development, National Institutes of Health.

Nourbakhsh, S., Mansor, M., Baba, M., & Madon, Z. (2013). The effects of multisensory method and cognitive skills training on perceptual performance and reading ability among dyslexic students in Tehran-Iran. *International Journal of Psychological Studies*, *5*(2), p92.

O'Connor, R. E. (2014). *Teaching word recognition: Effective strategies for students with learning difficulties*. Guilford Publications.

O'Connor, R. E., Jenkins, J. R., & Slocum, T. A. (1995). Transfer among phonological tasks in kindergarten: Essential instructional content. *Journal of Educational Psychology*, *87*(2), 202.

O'Connor, R. E., Notari-Syverson, A., & Vadasy, P. F. (1996). Ladders to literacy: The effects of teacher-led phonological activities for kindergarten children with and without disabilities. *Exceptional children*, *63*(1), 117-130.

O'Connor, R. E., Fulmer, D., Harty, K. R., & Bell, K. M. (2005). Layers of Reading Intervention in Kindergarten Through Third Grade Changes in Teaching and Student Outcomes. *Journal of Learning Disabilities*, *38*(5), 440-455.

O'Connor, R. E., & Vadasy, P. F. (Eds.). (2013). *Handbook of reading interventions*. Guilford Press.

Orton, S. T. (1925). Word-blindness in school children. *Archives of Neurology & Psychiatry*, *14*(5), 581-615.

Perfetti, C. A., Beck, I., Bell, L. C., & Hughes, C. (1987). Phonemic knowledge and learning to read are reciprocal: A longitudinal study of first grade children. *Merrill-Palmer Quarterly (1982-)*, 283-319.

Reschly, A. L. (2010). Reading and school completion: Critical connections and Matthew effects. *Reading & Writing Quarterly*, *26*(1), 67-90.

Ritchey, K. D., & Goeke, J. L. (2006). Orton-Gillingham and Orton-Gillingham—based reading instruction a review of the literature. *The Journal of Special Education*, *40*(3), 171-183.

Rose, T. E., & Zirkel, P. (2007). Orton-Gillingham Methodology for Students With Reading Disabilities 30 Years of Case Law. *The journal of special education*, *41*(3), 171-185.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score.*The American Statistician*, *39*(1), 33-38.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results.Psychological bulletin, 86(3), 638.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2003). Evaluation: A systematic approach. Sage publications.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology, 66(5), 688.

Rutter, M. (2006). Is Sure Start an effective preventive intervention?. *Child and Adolescent Mental Health*, *11*(3), 135-141.

Ryder, J. F., Tunmer, W. E., & Greaney, K. T. (2008). Explicit instruction in phonemic awareness and phonemically based decoding skills as an intervention strategy for struggling readers in whole language classrooms. *Reading and Writing*, *21*(4), 349-369.

Scarborough, H. S., & Brady, S. A. (2002). Toward a common terminology for talking about speech and reading: A glossary of the “phon” words and some related terms. *Journal of Literacy Research*, *34*(3), 299-336.

Scargle, J. D. (2000). Publication Bias: The “File-Drawer” Problem in Scientific Inference. Journal of Scientific Exploration, 14(1), 91-106.

Scheffel, D. L., Shaw, J. C., & Shaw, R. (2008). The efficacy of a supplementary multisensory reading program for first-grade students. *Journal of Reading Improvement*, *45*(3), 139-152.

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. Journal of Counseling psychology, 57(1), 1.

Schuele, C. M., & Boudreau, D. (2008). Phonological awareness intervention: Beyond the basics. *Language, Speech, and Hearing Services in Schools*,*39*(1), 3-20.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiology and drug safety, 17(6), 546-555.

Shadish, Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Wadsworth Cengage learning.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103*(484), 1334-1344.

Shadish, W. R., & Steiner, P. M. (2010). A primer on propensity score analysis.*Newborn and Infant Nursing Reviews*, *10*(1), 19-26.

Share, D. L., Jorm, A. F., Maclean, R., & Matthews, R. (1984). Sources of individual differences in reading acquisition. *Journal of educational Psychology*, *76*(6), 1309.

Simmons, D. C., Coyne, M. D., Kwok, O. M., McDonagh, S., Harn, B. A., & Kame'enui, E. J. (2008). Indexing response to intervention a longitudinal study of reading risk from kindergarten through third grade. *Journal of Learning Disabilities*, *41*(2), 158-173.

Slocum, T. A., O'Connor, R. E., & Jenkins, J. R. (1993). Transfer among phonological manipulation skills. *Journal of Educational Psychology*, *85*(4), 618.

Smith, J. W., & Elley, W. B. (1995). *Learning to read in New Zealand*. Richard C. Owen Publishers.

Snow, C. E., & Juel, C. (2005). Teaching children to read: What do we know about how to do it?.

Spear-Swerling (2013). Phases in Reading Words and Phonics Interventions. In O'Connor, R. E., & Vadasy, P. F. (Eds.). *Handbook of Reading Interventions* (pp.). Guilford Press.

Spybrook, J., Raudenbush, S. W., Liu, X. F., Congdon, R., & Martínez, A. (2006). Optimal design for longitudinal and multilevel research: Documentation for the "Optimal Design" software. Survey Research Center of the Institute of Social Research at University of Michigan.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading research quarterly*, 360-407.

Steiner, P. M., & Cook, D. (2013). Matching and propensity scores. *The Oxford Handbook of Quantitative Methods in Psychology*, *1*, 237.

Stevens, J. (2002). Applied multivariate statistics for the social sciences (4th ed.). Hillsdale, NJ: Erlbaum.

Stoner, J. C. (1991). Teaching at-risk students to read using specialized techniques in the regular classroom. *Reading and Writing*, *3*(1), 19-30.

Strauss, A. A., & Lehtinen, L. E. (1947). Psychopathology and education of the brain-injured child.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. Journal of Educational psychology, 51(6), 309.

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, *46*(1), 90-118.

Torgesen, J. K. (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research & Practice*, *15*(1), 55-64.

Torgesen, J. K. (2002). The prevention of reading difficulties. Journal of school psychology, 40(1), 7-26.

Torgesen, J. K., Morgan, S. T., & Davis, C. (1992). Effects of two types of phonological awareness training on word learning in kindergarten children. *Journal of Educational psychology*, *84*(3), 364.

Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities immediate and long-term outcomes from two instructional approaches. *Journal of learning disabilities*, *34*(1), 33-58.

Torgerson, C., Brooks, G., & Hall, J. (2006). *A Systematic Review of the Research Literature on the Use of Phonics in the Teaching of Reading and Spelling*. DfES Publications.

Trochim, W. M. (1984). Research design for program evaluation: The regression-discontinuity approach (Vol. 6). Sage Publications, Inc.

Tunmer, W. E., & Nesdale, A. R. (1985). Phonemic segmentation skill and beginning reading. *Journal of Educational Psychology*, *77*(4), 417.

Tunmer, W. E., & Arrow, A. W. (2013). Reading: Phonics instruction.*International guide to student achievement*, 316-319.

Tunmer, W. E., Chapman, J. W., Greaney, K. T., Prochnow, J. E., & Arrow, A. W. (2013). Why the New Zealand National Literacy Strategy has failed and what can be done about it: Evidence from the Progress in International Reading Literacy Study (PIRLS) 2011 and Reading Recovery monitoring reports. *Australian Journal of Learning Difficulties*, *18*(2), 139-180.

Uhry, J. K., & Clark, D. B. (2005). *Dyslexia: Theory & practice of instruction*. Pro Ed.

Van der Klaauw, W. (2002). Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression–Discontinuity Approach*. International Economic Review, 43(4), 1249-1287.

Van Hell, J. G., Bosnian, A. M., & Bartelings, M. C. (2003). Visual dictation improves the spelling performance of three groups of Dutch students with spelling disabilities. *Learning Disability Quarterly*, *26*(4), 239-255.

Vadasy, P. F., Jenkins, J. R., Antil, L. R., Wayne, S. K., & O'Connor, R. E. (1997). The effectiveness of one-to-one tutoring by community tutors for at-risk beginning readers. Learning Disability Quarterly, 20(2), 126–139.

Vadasy, P. F., Sanders, E. A., & Peyton, J. A. (2005). Relative effectiveness of reading practice or word-level instruction in supplemental tutoring: How text matters. Journal of Learning Disabilities, 38(4), 364–380.

Vaughn, S., Mathes, P., Linan-Thompson, S., Cirino, P., Carlson, C., Pollard-Durodola, S., Cardenas-Hagan, E., & Francis, D. (2006). Effectiveness of an English intervention for first-grade English language learners at risk for reading problems. Elementary School Journal, 107(2), 153–180.

Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Small, S. G., Pratt, A., Chen, R., & Denckla, M. B. (1996). Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. Journal of Educational Psychology, 88(4), 601.

Vellutino, F. R., Scanlon, D. M., Small, S., & Fanuele, D. P. (2006). Response to Intervention as a Vehicle for Distinguishing Between Children With and Without Reading Disabilities Evidence for the Role of Kindergarten and First-Grade Interventions. *Journal of learning disabilities*, *39*(2), 157-169.

Walberg, H. J., & Tsai, S. L. (1983). Matthew effects in education. *American Educational Research Journal*, *20*(3), 359-373.

Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2008). Investigating ceiling effects in longitudinal data analysis. Multivariate behavioral research, 43(3), 476-496.

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

What Works Clearinghouse (2012). What Works Clearinghouse procedures and standards handbook v. 2.1.

Wilkinson, I. A., Freebody, P., & Elkins, J. (2000). Reading research in Australia and aotearoa/new zealand.

Wise, B. W., Ring, J., & Olson, R. K. (1999). Training phonological awareness with and without explicit attention to articulation. *Journal of Experimental Child Psychology*, *72*(4), 271-304.

Wyse, D., & Styles, M. (2007). Synthetic phonics and the teaching of reading: the debate surrounding England's ‘Rose Report’. *Literacy*, *41*(1), 35-42.

Wyse, D., & Goswami, U. (2012). Early reading development. *Handbook of early childhood literacy*, 379-394.

Yeh, S. S., & Connell, D. B. (2008). Effects of rhyming, vocabulary and phonemic awareness instruction on phoneme awareness. *Journal of Research in Reading*, *31*(2), 243-256.

Young, R., & Johnson, D. R. (2010, May). ‘Imputing the Missing Y’s: Implications for Survey Producers and Survey Users. In Proceedings of the AAPOR Conference Abstracts (pp. 6242-6248).

Appendix A: OG Institute's Ten Step Lesson Plan

Teachers are instructed to progress through the ten lesson plans as designed. The program is structured so two lessons or *steps* are completed each day in the 20 minutes set aside for the OG Institute intervention. Depending on the lessons, time of the year, and needs of the students, this may vary slightly. After the teacher completes step (or lesson) ten, they begin again with step one, but with the next letter or topic on the scope and sequence chart (Appendix B).

Step 1: Phonemic Awareness

Kindergarten students begin developing phonemic awareness skills by first listening to *environmental sounds*. The students listen to a short recording containing one or more different sounds (animals, noises, etc). The pattern of the sounds will vary, sometimes with a sound repeating. First, the students are asked to count how many different sounds they hear. Second, using a different colored block for each sound they hear, the students are asked to sequence the sounds from left to right. The teacher is instructed to use key vocabulary terms such as: same, different, before, after, between, first, middle, next, last.

After students have become familiar with environmental sounds they will use Elkonin boxes with *finger spelling* in this step for the rest of the intervention. Students spend a few lessons on a single isolated phonemes before moving onto the activity *Words* *with Blocks*. In this activity the teacher says a word followed by the students repeating it. The teacher will then use the word in a sentence followed by the students repeating the entire sentence. The teacher will repeat the word once more and then *tosses it* (pretends to throw it) to the students. The students repeat the word as they pretend to catch the word in their non-dominant hand. The students represent each phoneme in the word one at time, saying it as they raise a finger with each phoneme. The students place blocks on their Elkonin box mat for each sound, using the same color for same sounds (e.g. the word dad could be represented by the pattern blue-red-blue)

Step 2: Three-Part Drill

The first part of the lesson involves using visual stimuli. The teacher holds up a flash card with a letter. The students then say the sound. The teacher puts each letter used on the *blending board* (an upright board with clear pockets to holds so the entire group can see them). The second part of the lesson involves auditory and tactile stimuli. The teacher says the sound but does not show the letter. The students say the sound then repeat the sound as they write the letter in a *sand tray* (Usually a paper plate filled with salt, so that the student's tracing pattern remains). The last step teaches blending using a blending board. The teacher randomly removes cards from the board. The students tap out the letters saying each sound. For example: /l/ - /o/ - /d/ then sweeps under the word and blends the word /lod/. After students have gained sufficient skills, the teacher eliminates saying the word aloud; instead, have students say them "in their heads". When signaled, the students will read the words as a whole.

Step 3: Vowel Intensive

The lesson is composed of four sections that introduce students to vowels. Each of the four sections increases in difficulty. The teacher moves to the next section after students have mastered the first. The first section only includes isolated vowel sounds. The teacher says the isolated vowel sounds (if needed the teacher will cue with body language to help them form the sound in their mouths). The students then repeat the vowel sound. The student will hold up the corresponding card and say the sound (for example /a/). The teacher repeats for all short vowels a,e,i,o,u. In the next section the teacher uses words of the form VC, using the same consonant with each vowel (e.g. ag, eg, ig, og, ug). The teacher then moves to using the same form but with different consonants after the vowels (e.g. ap, ab; en, et; im, id; oc, oj; uf, un). For the last section of lessons the teacher uses words of the form CVC which all consist of nonsense words (e.g. lat, zan, zeg, med, hib, fid, etc).

Step 4: Introducing a New Concept A: Letter Formation

This step introduces students to letters and letter combinations. The teacher introduces a new letter with a letter card and object/picture that uses the letter. The student then finds the letter on their *alphabet strip* (a long strip of paper containing all 26 letters) & classify it as a vowel or consonant. The students follow a series of repetitive steps to practice writing the letter. A graphical 'house' (Figure A.1) is used aid in the writing of the letter. It has an attic and a basement used to convey the how letters can have different heights. The teacher introduces the strokes using a large house diagram. The student will then *skywrite* (writing in the air) while saying the *strokes* three times. Each stroke has a name associated with it that become familiar vocabulary with the students. Next, the student traces, while saying the correct names of the strokes, the letter on a handout with the letter already drawn with hyphenated line. The student repeats this three times. The teacher displays the letter on House Paper and discusses the letter placement. While saying the names of the stroke out loud, students will do one each of the following four activities: 1) Tracing the preprinted letter on the house paper with their index finger three times, 2) Tracing the letter using a green crayon with a screen underneath three times 3) Tracing the crayon bumps with the index finger three times and 4) Places the screen on top of the paper and traces with their index finger three times. The student then writes the letter on paper with green crayon without the screen in three steps. First, on dotted letter, next, with only a dot as a starting point and finally, without scaffolds. Intermittently between the steps above the teacher will insert auditory blending practice. Starting with the thumb, the teacher holds up one finger and says each phoneme in the word. When the teacher puts all of the fingers together, the students say the word. The teacher asks the students to point to the finger representing the sound.

Step 5: Introducing a New Concept B: Sound-Symbol

In this step students continue to practice letters by rotating through several activities centered on multisensory stimuli. In each activity the students trace the letter while saying the sound. The activities include using the stroke handout, skywriting (the letter must be displayed in a pocket chart or the blending board so that students can trace the letter in the air), a screen on top of letter card, glue letter, felt paper or the sand tray.

Step 6: Introducing a New Concept C: Making Connections to Prior Knowledge

Students are asked to brainstorm words containing the given sound and the teacher records them on a chart. For students with limited vocabulary or poor word retrieval, students are provided clues (e.g. the teacher would say, "I am thinking of something you write with. It has an eraser. It is a…pencil"). The teacher highlights the new concept on the phoneme/grapheme chart and adds the card to the review deck for use during the next Three-Part Drill.

Step 7: Spelling and Blending with Letter Tiles

In this step students will spell words using *letter trays* (these trays hold the student flash cards in an upright fashion, so the students can add and remove cards to build words). The trays should include only letters that have been taught directly during intervention. The teacher says a word containing a newly taught sound. The student then repeats the word. The teacher then uses the word in a sentence and the students repeat the sentence. The teacher, says the word again and "tosses" the word to the students. The students say the word as they catch it in their non-dominant hand. Starting with the thumb, the students represent each phoneme with a finger as they say the phonemes. The students then they select letters from the available letters on the letter tray to build the word. When directed by the teacher, the students tap each letter, say the sounds, and blend them together to form the word

Step 8: Dictation

In this step students will do one of three activities. The first is to practice spelling using *dictation paper* (lined paper with several lines for students to practice writing). The teacher says a word that includes only known letters and the student repeats it. The teacher then says a sentence with the word and the students repeat the sentence. The teacher again says the word and tosses the word to the students. As in previous steps the students catch the word in their non-dominant hand while repeating the word. The students will first fingerspell saying sounds before they write the word with a pen. After the students have written the word on dictation paper the teacher provides correct spelling. The students check for accuracy and correct their work by striking through and correcting errors. The students then rewrite the words in the second column of the dictation paper. The second activity practices fluency. The teacher points to words on a list and the students read them chorally. The teacher will read from the list in different directions changing from top to bottom, bottom to top and randomly. In the third activity students will practice spelling entire sentences using words that include letters and words that they know. The teacher begins this activity by saying a sentence. The students are instructed to visualize the sentence, only repeating the words in their mind. The teacher repeats the sentence while pounding their fist on the table with the rhythm of the sentence. The students repeat the sentence out loud, repeating the rhythm with their fists. The students then write the sentence on lined paper. The teacher then provides the correct sentence to the students. The students then check their work for accuracy, striking through and correcting errors. The students then rewrite the sentence. The students are then instructed to write their own sentences, underlining all red words with a red crayon.

Step 9: Red Words (either a new red word or review)

Red Words are how the OG Institute refers to common, but phonemically irregular words (like *the*) that students must recognize quickly without phonetic decoding. If the students are ready for a new word the teacher will select a new red word as indicated in the Scope & Sequence chart (Appendix B). The teacher presents a red word printed on a red on a flash card. The teacher will pronounce the word, define it, and use it in a meaningful sentence. The teacher then distributes the red word the students on a piece of paper. While tracing the word with index finger, the students say the letters and sweep under the word saying the word three times. The student repeats this with screen underneath the red word paper. The students say the letters again while tracing the word with a red crayon, and then sweep under the word, saying the word three times. The students are then instructed to stand. While looking at the word displayed in the pocket chart the students will rhythmically tap down their arm saying the letter names three times (e.g. "t-h---e-y → /they/"), then making a sweeping motion on their arm. The students will then trace the crayon bumps, made in the previous step, with their finger while saying the letters. The follow this with a sweep under the word saying the word three times. Next, with a screen on top of the paper, the students will trace the letters with their index finger and say the letter names followed by a sweep under the word, saying the word three times. Following this, on the back of the paper, the students will write the word from memory with a red crayon and the teacher checks the student's work for accuracy. Lastly, the student writes the word two more times with a red crayon. The student is then given a sentence with a blank. The student reads the sentence, writes the red word in the blank and then rereads the sentence.

If no new word is indicated, the teacher will review previously learned words. One at a time, the teacher displays the red word cards in the pocket chart. The students will stretch out their non-dominant arm. Using their dominant hand, the students will tap down their outstretched arm, moving from the shoulder down to the wrist. The students rhythmically tap one time for each letter as they say the letters aloud (e.g. /co…me/). The students then sweep the entire arm (from shoulder to wrist) and say the word (e.g. /come/). If additional practice is required, the students will place the screen over the words in their red word books. They should trace the words with their index fingers while saying the letter names.

Step 10: Reading Fluency

Students will read the next selection according to the Scope and Sequence Chart (Appendix B). If no new story is indicated the students will read previously read stories.

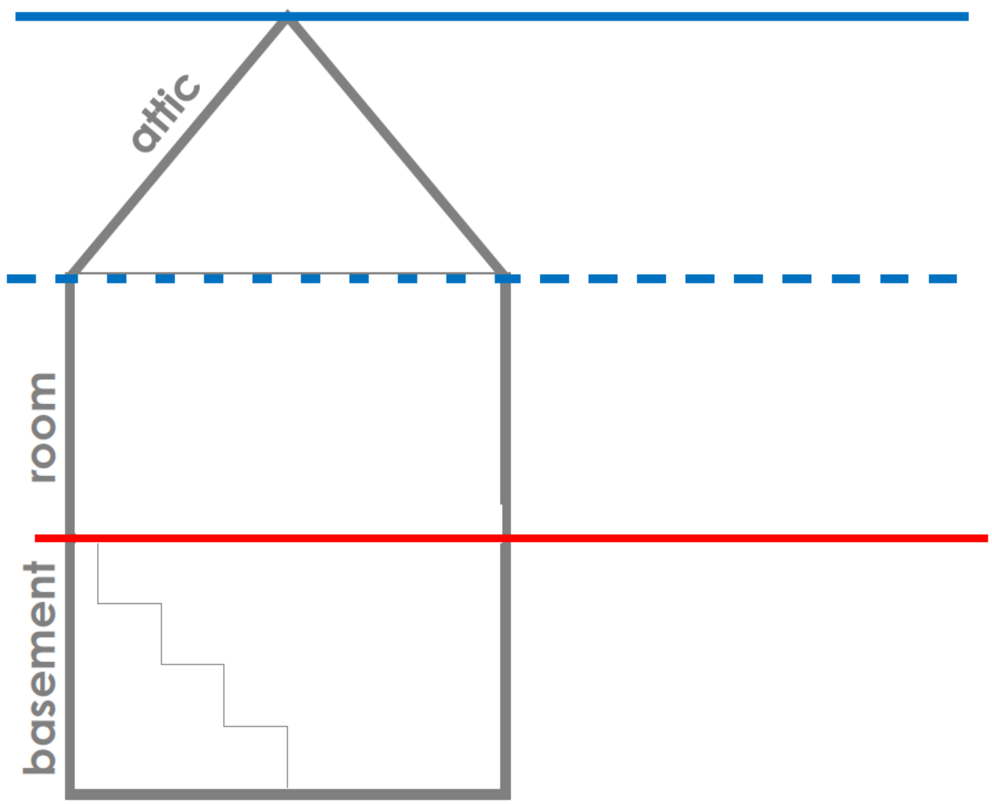

Figure A.1: A 'House' Diagram used to help students with letter formation when learning to write a new letter

Appendix B: Scope and Sequence of Topics for the OG Institute (Kindergarten)

	Letters	Card #	Bonnie Kline Red Words	Bonnie Kline Story Title	Booklet
Unit 1 Kinder	short a, t, m, s	#1-4	I [1]	Tat	1
Unit 1	b	#5		Matt, Sam	1
Unit 1	c	#6	a [2]	Cam	1
Unit 1	f	#7	the [3]	Fat Cat; The	1
Unit 2	h	#8		Sam's; Mac's Cat	2
Unit 2	r	#9		Tat	2
Unit 2	j	#10		Tam	2
Unit 2	n	#11		Fan	2
Unit 2	p	#12		Pam At Bat; Fat Sam; Tan	2
Unit 2	l	#13	has [4]	Hal, Cal, and Sal; Pal's Ran; Has	3
Unit 3	g	#14		Jan's Bag; Tag, the Cat	3/4
Unit 3	d	#15		Jabs; Tad's Ham; Bad Rags	4
Unit 3	v	#16	to [5], of [9]	Val's Van; Dad's Van; To; Tag Dan	5
Unit 3				A Lot of	5
Unit 4	short i	#17		Sig; Tim's Hat	6
Unit 4	w	#18	have [6]	Nan's Wig; Bad Mac; Have	6
Unit 5-6	k	#19		Kim	7
Unit 5-6	-ck	#20	said [7]	Kick the Can; Kim's Cap; Said	7
Unit 7	qu	#21		Quack	8
Unit 7	x	#22		Fix the Cab	8
Unit 7	y	#23		Max	8
Unit 7	z	#24		Zap and Zig-Zag	8
Unit 8	short o	#25		On Top	9
Unit 8	-ff	#26	was [8]	A Hotdog and Pop; A Bad Dog?	9
Unit 9	-ll, -ss, -zz	#27-#29	you [10]	Bill's Mill; Don's Bass; A Kiss; Dad's Pop; You	10
Unit 10	-ng*, -nk*	#30-37	what [11]	Rock King Gang; At the Rink; Bad King; What	11
Unit 11s Grade 1	u	#68	love [4a]	Is It a Dog?; But a Pup; Love; Bad Luck	11s
Unit 11	wh*	#38	are [11a]	Hank's Big Cats	12
Unit 11	ch*	#39	ball [12]	Chad's Ax; Ping-Pong	12
Unit 11	th*	#40		Math; Whiff-Whiff	12
Unit 11	sh*	#41	put [13]	A Pink Dish; A Quiz; The Path	13
Unit 12	short e	#42	could [14] should [15] would [16]	Ed; The Shell Shop;Should, Could, Would; Doc Mel; Ten Hens	14
Unit 13	sl; bl	#43-44		Ned's Sled; Black Blob	15
Unit 13	fl, cl	#45-46		Black Blob;	15
Unit 13	gl, pl	#47-48	they [17]	A Glass Dish; Lots of Things	15
Unit 13	br, cr	#49-50		Crabs	16
Unit 13	fr, dr	#51-52		Fred	16

superscript = red word card number (eye [48])
*Kindergarten may skip

	Letters	Card #	Bonnie Kline Red Words	Bonnie Kline Story Title	Booklet
Unit 13	tr, gr, pr	#53-55		Trash; Trish's Doll	16
Unit 13	thr, shr	#56-57	your[18]	Mom's Dress; Your Mess	16
Unit 13	sc, sm	#58-59		A Bad hit	17
Unit 13	sl, sn	#43, 60		The Box	17
Unit 13	sk	#61		Scott's Red Sled	17
Unit 13	sp, st; squ	#62-64		Stan's Stick; Squids	17
Unit 13	sw, tw, dw	#65-67		Ring-a-Ling; The Shrill Yell; The Twins; Bill's Frog	18
Unit 14	short u review	#68		Mud; Brad's Hut; Chuck's Drum; A Bad Trip?; Mom's Muffins	19
Unit 15	closed syllables			Justin's Log Cabin; Dan D. Duck; Fantastic Pocket; What Pals	20
Unit 16	spr	#69		Springs	21
Unit 16	str	#70	were[19]	Wet Socks; Black Angus Ranch	21
Unit 16	scr, spl	#71-72	two[20]	Spring; Two	21
Unit 17	-nt	#73		A Stunt	22
Unit 17	-st	#74		Hot Cross Buns	22
Unit 17	-sk, sp	#75-76	do[21]	Ask Mr. Best; You and I; It's a Musk Hog	22
Unit 17	-nd, -mp	#77-78	many[22] any[23]	Ed's Brass Band; Song of the Swamp; Many Things; With Any Luck	22
Unit 17	-ld, -lk,-lp, -lt	#79-82		A Duck's Lunch	23
Unit 17	-ft, -ct, -pt, -nch	#83-86		A Grand Lunch	23
Unit 18 Grade 2	**-tch**	**#87**	**such[24] much[25] rich[26] which[27]**	**Catnap at the Gift Shop**	**24**
Unit 19	ar	#88	who[28]	Sharks; I Wish; Who	25
Unit 19	er, ir, ur	#89-#91	come[29] some[30]	The Perfect Supper; A Big Run; Fur-Fur; Come; Red Bird; Junk	25
Unit 19	or	#92		Storks	26
Unit 20	open syllable	#1, 17, 25, 42, 68	whose[31] once[32]	Ho-Bo Hux; Whose?; Just Once	27
Unit 20-21	a-e	#93	very[33]	A Big Problem; Chicken Pox?; Robot; Fishing at Sho-Lo Park	28
Unit 21	i-e	#94		The Bike Trip	28
Unit 21	o-e, e-e, u-e	#95-97	here[34]	A Joke on Pete; A Close Game; A Swell Spot	28
Unit 21			live[35] give[36] from[37]	Wetlands; Slither! Slither!; A Gift; Eggs; A Problem	29
Unit 22	vccv, vcv			Rover; Insect Charmer; Otto Wingate's Band	30
Unit 23	y as a vowel	#98	where[38] there[39]	The Sly Fox; Where?; There	31
Unit 24a	ai	#100		Mail for the Jail; A Bad Storm	32
Unit 24a	-ay	#101	whom[40]	Rays; Whom	32
Unit 24b	ea, ee	#102-103	great[46]	Sammy Snake; It's Great!	33

superscript = red word card number (eye [48])

*Kindergarten may skip

	Letters	Card #	Bonnie Kline Red Words	Bonnie Kline Story Title	Booklet
Unit 24b	-ie (long e), -ey	#106		A Happy Valley	33
Unit 24c	-ie (long i)	#106	does [41]	A Pie, a Tie, and a Lie; Does	33
Unit 24d	oa	#107		Off To the Park	34
Unit 24d	-ow	#108		It May Snow Tomorrow	34
Unit 24d	-oe	#109	other[47] mother[49] brother [51]	Joe's Garden; My Other Brother	34
Unit 25	ea (short)	#102		What's the Weather?	34
Unit 26	schwa			Rascal	35
Unit 27	-ble, -dle, -gle, -kle, -ple, -tle (-cle, -fle, -zle)	#110-118	done [42] one [43]	King Clam; It's Done; One Math Lesson	35
Unit 28	-dge	#119	father [50]	Hodgepodge, My Father	36
Unit 29a	oo (moon)	#120	eye [48]	Fantastic Grass; Eye	36
Unit 29b	oo (cook)	#120		A Good Thing	36
Unit 30	oi, -oy	#121-122	watch [52]	Butter, Corn, and Soap; Watch It!	37
Unit 31	-ow	#108		There's a Clown in Town	37
Unit 31	ou	#123		The Land of Loons	37
Unit 33	soft c	#124	friend [60]	Ice Fishing; Best Friends	38
Unit 33	soft g	#125	again [58]	It Looks Like a Giraffe; Again	38
Unit 34	au, -aw	#126-127		Power Swooper	39
Unit 34	augh	#128	against [59]	A Dog and a Mom; Against	39
Unit 34	-ew	#129		Trail to Sunset Ridge	39
Unit 34	-ue	#130		Blue Sue	39
Unit 35	ui	#131	beautiful [65]	Just Pretend!	40
Unit 35	oo graphemes		been [66]	Bandit	40
Unit 35			blood [67]	It Has Blood In It	BK #2
Unit 35	seven syllables		none [68]	A Problem at Cripple Creek; None	BK #2
Unit 35			only [69] says [70] sure [71]	Only; Says Who; Sure	BK #2

Appendix C: Fidelity Checklist for the OG Institute (Kindergarten)

Instructional Component	Evidence of Implementation	No Evidence
STEP 1: Phonemic Awareness. **K:** Students represent 1-3 isolated phonemes or 2-3 sound words with blocks in Elkonin boxes. **G1:** Students represent 2-4 sound words with blocks or compare two similar words containing 2-4 sounds and make block changes where the words differ (Chaining.)		
STEP 2: Three-Part Drill **Visual:** The teacher flashes cards and students say the sounds. **Auditory:** The teacher says sounds and students repeat and write them in the salt/sand tray. **Blending Board:** Teacher removes one card at a time. Students blend the phonemes and say the word chorally. (Student may blend silently and respond with the entire word chorally.)		
STEP 3: Vowel Intensive **Version 1:** Teacher says vowel sound, students repeat and hold up letter tent or card. All four steps outlined below are followed. A. Vowel in isolation B. VC with the same consonant following the vowel C. VC with different consonants following the vowel D. VCV in nonsense words **Version 2:** Teacher says word containing a vowel that can be spelled in more than one way. Students repeat word and hold up card containing the correct spelling for the word.		
STEP 4: Introduce a New Concept 1. Displays letter in pocket chart or blending board. Associates it with a picture or an object. 2. Student identifies target letter on alphabet strip and classifies letter as a vowel or consonant. 3. Teacher introduces the letter strokes for drawing the letter. (K and beginning of 1st grade only.) a. Students skywrite the letter saying the strokes 3x. b. Students trace stroke paper saying the strokes 3x. c. Teacher discusses letter location using house paper. d. On house paper, students trace letter with finger and say strokes 3x. e. Students trace letter with green crayon and screen underneath paper with saying the strokes 3x) f. Students use finger to trace crayon marks while saying the strokes 3x. g. Students place screen on top of house paper, trace the letter, and say the strokes 3x. h. Students trace the dotted letters on house paper with green crayon while saying strokes. i. Students form the letter independently while saying strokes.		

STEP 5: Sound-Symbol Relationship VAKT: Students say the sound while tracing the letter using multisensory techniques: Stroke handout, skywrite, glue letter, felt paper, salt/sand tray, screen on top of a letter card, etc.		
STEP 6: Making Connections Students brainstorm words containing the targeted letter(s) at the beginning, middle, and end position of words, as appropriate. The teacher lists the words and may give clues for words as a scaffold.		
STEP 7: Spelling Students spell words using letter tiles (K-1) or white boards (G2). 1. Teacher says word; students repeat. 2. Teacher says sentence; students repeat. 3. Teacher says word while "tossing it"; students say word while "catching it". 4. Students release words putting up one finger per sound as they say the phonemes (starting with the thumb.) 5. Students build word with letter tiles or write on a white board. 6. When directed, students tap each letter, say the sounds, and blend the word together.		
STEP 8: Dictation Using the same routine as above, students use pens to write words on dictation paper. Scaffolds should be removed, as appropriate. 1. Students write words in column 1, check and correct errors, and rewrite the words in column 2. 2. As directed by the teachers, students read the words in the second column (top to bottom, bottom to top, and randomly.) 3. Students write, correct, rewrite a sentence to dictation. 4. Students write their own sentence using a word in column 2.		
STEP 9: **Version A: New Red Word** 1. Teacher says the word, defines it, and uses it in sentence. 2. Students trace words & saying the letters and word 3x. a. With screen underneath trace with red crayon. b. Students trace crayon bumps with finger. c. With screen on top of the paper, students trace letters. 3. Tap down arm saying the letters and sweep arm saying the word. 4. Write word from memory 3x. 5. Completes final task (K-1 cloze; 2 alphabetize). **Version B: Review Red Words** Students tap letter names on arm and blend into a word.		
STEP 10: Reading Fluency Students read chorally or independently phonically-controlled passages that contain the targeted sound/symbols.		
Daily Repeated Reading Practice Students practice reading the most recently introduced passage until a new story is taught.		
Progress Monitoring Teachers progress monitor students every 2-3 weeks using DIBELS to assess reading growth.		