Application of the Regression Discontinuity Technique to the Response to
Intervention (RTI) Model of Service Delivery for Determining the Effects of
Early Intervention in Reading

Dissertation

Submitted in Partial Fulfillment of Requirements
For the Degree Doctor of Philosophy in the Graduate School of The Ohio State
University

By
Shannon Kathleen Flaum

Graduate Program in Education
The Ohio State University
2009

Dissertation Committee
Dr. Dorinda Gallant, Co-Advisor
Dr. Ann A. O’Connell, Co-Advisor
Dr. Laurice Joseph, Committee Member
ABSTRACT

The recent reauthorization of the Individuals with Disabilities Education Act (IDEA) into the Individuals with Disabilities Education Improvement Act (IDEIA) has had a profound impact on the manner in which schools provide services to their students. The reauthorization, coupled with the adoption of the No Child Left Behind Act (NCLB), emphasizes needed improvements in the general education and special education systems and requires that curricula and instructional tools demonstrate proven effectiveness. These recommended improvements have led to the birth of the Response to Intervention (RTI) model of service delivery.

The purpose of this study was to apply the regression discontinuity design to determine the effectiveness of the RTI model of service delivery for a group of first-grade students receiving Tier 2 intervention services in reading. An additional goal of this study was to determine the appropriateness of regression discontinuity as a method of analysis for the RTI model.

The independent variable for this study was participation in Tier 2 reading services. The dependent variable was the students’ post-program level of reading fluency, as measured by their performance on the spring administration of the Oral Reading Fluency (ORF) subtest of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). Students’ performance on the pre-program measure, the winter
administration of the Oral Reading Fluency subtest of the DIBELS, was used to assign students to Tier 2 services. Out of 141 students, 21 met criteria for Tier 2 services and participated in the program group; the remaining 120 students participated in the control group.

Results of this study yielded standardized effect sizes of .176 and .128 for two models, suggesting that Tier 2 services were effective in increasing reading fluency for this group of first-grade students. The strengths and limitations of using the regression discontinuity design are discussed. The findings in this study added to current research on the effectiveness of Tier 2 services and appropriateness of using the regression discontinuity design.
Dedication

Dedicated to my beloved grandparentes, Edna and Beecher Rogers, and sons, Kellen and Reilly
Acknowledgments

A project of this undertaking cannot be completed in isolation, and I would be remiss if I did not acknowledge the tremendous support I have received from my advisors, committee members, family and friends. To my advisor, Dr. Dorinda Gallant, I wish to express my sincere gratitude for your mentorship, knowledge, critical eye, and belief in my ability to complete such a project. You are a Godsend. To my co-advisor, Dr. Ann O’Connell, I thank you so much for introducing me to the challenging world of multiple and hierarchical linear regression and for your support and feedback throughout this process. To Dr. Laurice Joseph, I thank you for your knowledge of RTI, your willingness to serve on my committee, and for making school psychology fun again. These intelligent, talented women have challenged me and in each their own way, provided me with knowledge and experiences that I will carry with me throughout the rest of my academic career. To Dr. Richard Lomax, I sincerely express my thanks for introducing me to structural equation modeling and agreeing to serve on my candidacy committee. Your tangential comment about regression discontinuity, made during my candidacy examination, was the seed for this dissertation, and I am thankful for it. To Jack Stephens, school psychologist of Mad River Local Schools, I express my sincere gratitude in your willingness to take time from your very busy schedule to provide me with data and patiently answer my questions. This project would not exist without your help and support. To Dr. James Demery, Director of Research at McGraw-Hill, I cannot
express how much I value your friendship and mentorship. Your encouraging phone call the morning of my candidacy exam, and support and guidance throughout the last several years have meant so much to me. To my ‘quantitative research’ friends, Melinda, Terri, Sandy, and Tiffany, thanks for the camaraderie and friendship. Melinda, I will carry happy memories of log linear with me! To my late grandparents, Edna and Beecher Rogers, you are the definitions of hard work, commitment, and perseverance. During the many peaks and valleys, thoughts of you provided that extra push of motivation. I love and miss you both. To my parents, Barbara and Chuck Andrews and Steve and Carol Flaum, and grandparents, Joan and Chuck Flaum, thank you all for your support and understanding as I completed this crazy project. I would not be where I am today without your love and support. To my husband, Eric, completing this would not have been possible without your love and encouragement. I appreciate you more than you realize, and I thank you for believing in me. Finally, to my sons, Kellen and Reilly, I’m finished...mommy is yours!
VITA

June 1994…………………………………… B. S., Psychology and Biology
Wright State University

June 1999…………………………………… M.A. + 30, School Psychology
The Ohio State University

July 2002 to present………………….School Psychologist, Pickaway
County Educational Service Center

FIELDS OF STUDY

Major Field:  Education

Minor Field: Quantitative Research, Evaluation, and Measurement
# Table of Contents

Abstract ........................................................................................................................................ ii

Dedication ....................................................................................................................................... iv

Acknowledgments ............................................................................................................................ v

Vita .................................................................................................................................................... vii

List of Tables ..................................................................................................................................... xi

List of Figures ...................................................................................................................................... xii

Chapter 1 Introduction ..................................................................................................................... 1

  Background of the Study ................................................................................................................ 2

  Purpose of the Study ....................................................................................................................... 5

  Significance of the Study ................................................................................................................. 6

  Potential Implications of the Study ................................................................................................. 6

  Organization of the Dissertation .................................................................................................... 7

Chapter 2 Review of Literature ......................................................................................................... 8

  History of Special Education ......................................................................................................... 8

  Key Court Cases .............................................................................................................................. 10

  IQ-Discrepancy Model ................................................................................................................... 14

  History of RTI ................................................................................................................................. 16

  Overview of RTI .............................................................................................................................. 18
The 3-Tiered Model .................................................................20
Potential Benefits of RTI .............................................................24
Criticisms of RTI .................................................................26
Research on Effectiveness of Tier 2 ............................................27
Relevant Research Using Regression Discontinuity ......................31
Chapter 3 Methods ..................................................................33
Source of Data .........................................................................33
Measures for Pre- and Post-Assessment ......................................34
Assignment of Participants .........................................................38
Tier 2 Intervention Program .......................................................41
Reading Intervention for Early Success (Early Success) ..................42
Word Boxes .............................................................................43
Research Design .......................................................................44
Overview of Regression Discontinuity ........................................45
Logic of Regression Discontinuity ..............................................47
Regression Discontinuity and True Experiments .........................51
Assumptions of Regression Discontinuity ....................................52
Internal Validity ........................................................................54
Power and Effect Size ...............................................................56
The Statistical Model ................................................................59
Description of Program Fidelity ..................................................61
Chapter 4 Results ....................................................................63
Steps for Conducting Regression Discontinuity ................................................................. 63

Results for Regression Discontinuity .................................................................................... 68

Qualitative Information ........................................................................................................ 81

Chapter 5 Discussion ............................................................................................................ 84

Summary of Findings ............................................................................................................ 84

Effects of Tier 2 Instruction in Reading ................................................................................. 84

Utility of Regression Discontinuity ....................................................................................... 89

Limitations of the Study ....................................................................................................... 91

Suggestions for Further Research ......................................................................................... 93

Conclusions ......................................................................................................................... 93

References .......................................................................................................................... 95
List of Tables

Table 3.1 Benchmark Scores Corresponding to Classification of Risk on the DIBLES..........................................................................................................................40

Table 3.2 Student Gender by Group ..............................................................................................................40

Table 4.1 Descriptive Statistics For First-Grade Winter and Spring ORF .........................69

Table 4.2 Descriptive Statistics For Log-Transformed First-Grade ORF Scores ..........71

Table 4.3 Skew and Kurtosis Comparisons .................................................................................................72

Table 4.4 Codes for Data Analysis .................................................................................................................73

Table 4.5 Regression Results for Model 1 (Full Quadratic Model) ............................................75

Table 4.6 Regression Results for Model 2 (Without Quadratic Interaction Term) ........76

Table 4.7 Regression Results for Model 3 (Without Quadratic Term) .......................................78

Table 4.8 Regression Results for Model 4 (Without Linear Interaction Term) ................79

Table 4.9 Benchmark Scores Corresponding to Classification of Risk, by Administration ................................................................................................................81
List of Figures

Figure 2.1 Response to Intervention Model of Service Delivery ........................................21
Figure 3.1 Hypothetical Data Demonstrating No Treatment Effect ....................................48
Figure 3.2 Hypothetical Data Demonstrating a Treatment Effect ........................................50
Figure 3.3 Questions Assessing Program Fidelity ................................................................61
Figure 4.1 Bivariate Distribution of Spring ORF and Assignment Variable .........................74
Figure 4.2 Bivariate Distribution of Spring ORF and Assignment Variable, Quadratic ....77
Chapter 1
Introduction

The recent reauthorization of the Individuals with Disabilities Education Act (IDEA) into the Individuals with Disabilities Education Improvement Act (IDEIA) has had a profound impact on the manner in which schools provide services to their students. This reauthorization, coupled with the adoption of the No Child Left Behind Act (NCLB), emphasizes needed improvements in the general education and special education systems and requires that curricula and instructional tools demonstrate proven effectiveness. These recommended improvements have led to the birth of the Response to Intervention (RTI) model of service delivery. RTI generally refers to a school-wide system of service delivery that addresses these needed improvements highlighted in IDEIA and NCLB, with the overarching goal of increasing successful academic outcomes for all students, and most notably, for those who are struggling academically or behaviorally. While the term response to intervention is not specifically included in the 2004 reauthorization of IDEIA, the law does state that school districts may use “a process which determines if a child responds to a scientific, research-based intervention as part of its evaluation procedure” (Sec. 614(b)(6)). Originally stemming from the perspective that the traditional IQ-discrepancy formula fails in its attempt to validly identify those students who are
learning disabled, RTI is now being adopted as a school-wide model of instructional delivery as well as a system for identifying students with special needs.

**Background of the Study**

The data for 2007 indicated that 2,563,665 US students age 6 through 21 were identified as learning disabled and received special education services under IDEIA, Part B (U.S. Department of Education Office of Special Programs, OMB#18200043), and a recent survey reveals that more states are adopting response to intervention models to meet the needs of these struggling learners. To gauge the perception and level of adoption of RTI practices, a national survey of special education state department directors was conducted in August of 2008 through the Special Education Leadership and Quality Teacher Initiative, BUENO Center-School of Education (Hoover, Baca, Wexler-Love, & Saenz, 2008). Survey results revealed that of the 86% of state directors who responded to the survey, all reported that RTI practices were either currently adopted or in development. Over one-third of respondents reported that they were planning to adopt RTI either in supplement or as replacement to the traditional IQ-discrepancy model. The survey’s implications, as suggested by the authors, are clear: large-scale, national RTI implementation is already underway, and “appears to be an eventual certainty in the United States” (p. 11).

Special education programs are costly, and allocated funds for special needs services are expected to increase. During the fiscal year 2008, 19.2 billion dollars were spent on special education funding, with appropriated funds for the 2009 and 2010 fiscal
years set at 21.5 and 23.8 billion dollars, respectively (Eggert, 2005). These additional funds are earmarked for children not currently qualifying for special education services, but who may require academic or behavioral interventions to succeed in the general education setting. To assist schools in the implementation of intervention activities, the federal government currently permits school districts to divert 15% of its special education budget into the implementation of intervention programs designed to address academic difficulties in the early grades, and last year, expanded these funds by 14 million dollars (Texas Educational Agency, 2008). The RTI model of service delivery is increasingly utilized by school districts to address a child’s particular problem immediately, before remediation becomes too difficult or costly. While many school districts are quick to adopt the RTI model, little research has been conducted to determine the overall effectiveness of this new model of service delivery, as it is practiced in schools. Social programs that, in theory, seem promising are sometimes adopted and implemented without prior demonstration of effectiveness. Failure to adopt programs without evidence of effectiveness is not new, as Campbell addressed in his 1969 article, *Reforms as Experiments*:

The United States and other modern nations should be ready for an experimental approach to social reform, an approach in which we try out new programs designed to cure specific social problems, in which we learn whether or not these programs are effective, and in which we retain, imitate, modify, or discard them on the basis of apparent effectiveness on the multiple criteria available…So long have we had good intentions in this regard that many may feel we are already at this stage, that we already are continuing or discontinuing programs on the basis of assessed effectiveness. It is the theme of this article that this is not at all so, that most ameliorative programs end up with no interpretable evaluation (p. 409).
While we have certainly made substantive progress in the research of our social programs since the time of Campbell’s 1969 article, we are often quick to assume a program’s effectiveness based on the efficacy of individual components rather than the process as a whole. The current conceptualization of RTI is that the model refers to a particular criterion for decision-making, rather than a particular set of procedures; in fact, RTI consists of an integrated set of tools, procedures and decisions. Much of the current research has focused on the efficacy of individual components of RTI, such as the implementation of particular interventions for addressing particular academic weaknesses, rather than on the process as a whole. In theory, if the individual pieces are effective, then the overall process is effective (VanDerHeyden, Witt, and Gilbertson, 2007). VanDerHeyden, Witt, and Gilbertson (2007) state:

There are at least two problems with the research thus far conducted in support of RTI models. First, implementing RTI means implementing an integrated set of procedures and components while correctly applying sequenced decision rules. The research conducted to date with few exceptions has focused primarily on the efficacy of the components individually but not on the efficacy of the RTI process as an integrated whole. In theory, if the components are effective, then the overall process would be expected to produce results; however, the question of whether the overall process is effective must also be addressed. The second issue is that most of the research has been conducted by well-funded research centers. At least for the intervention component, evidence-based interventions can markedly decrease the need for special education services when implemented by high integrity by a research associate who is paid to do that job (p. 226).

The importance of conducting research on RTI as implemented in ‘real world’, practical settings is further underscored in Burns, Appleton, and Stehouwer (2005):

…a similar comparison of the effectiveness of RTI models used in research to those that exist in practice seems necessary to determine if the effectiveness of a model differs according to the setting in which it is used and to determine in which setting the models are most effective’ (p. 383).
Commonly, research on RTI focuses on the effectiveness of a particular intervention for a particular student or small group of students with similar deficiencies, and often utilizes single-subject research designs and methods of data analysis. While particularly powerful in helping to determine the effect of treatment for an individual, single-subject and small group approaches may not provide justification for the overall use of RTI as a model of service delivery. Furthermore, although RTI appears to demonstrate promise, Hoover, Baca, Wexler-Love, and Saenz (2008) indicated that “additional research is needed to evaluate various aspects associated with this growing practice” (p. 2). The regression discontinuity design, discussed in the literature for 40 years but rarely used in education, has been proposed by several researchers as a ‘perfect match’ for investigating the effectiveness of various RTI approaches and may be useful as a method of investigating effectiveness (Gersten & Dimino, 2006). For reasons cited later in this study, the regression discontinuity design appears to lend itself to the analysis of the RTI model of service delivery.

**Purpose of the Study**

The purpose of this study was to apply the regression discontinuity technique to determine the effectiveness of the RTI model of service delivery for a group of students receiving Tier 2 intervention services in reading. Regression approaches were used to compare the academic performance of those students selected to receive Tier 2 intervention services in reading to those who did not receive tier 2 intervention services. The research questions of interest are as follows:
1. To what extent are early intervention services in reading effective for Tier 2 students, under an RTI model of service delivery?

2. What are the statistical benefits and short-comings in the regression discontinuity technique in studying the RTI model?

**Significance of the Study**

A review of the literature reveals that most of the current research on RTI has focused on individual components, or selected interventions investigated using single-subject or small group designs. This study seeks to focus on the effectiveness of the RTI model, as students proceed from Tier 1 to Tier 2 within the RTI framework. One of the promises of RTI is that it effectively and immediately addresses the needs of students, with a related goal of reducing the number of students placed in special education programs. An understanding of the effectiveness of Tier 2 services will provide valuable information about the RTI as a model of service delivery.

**Potential Implications of the Study**

As a primary goal, it is hoped that the results of this study can assist policy makers and school districts in making informed decisions about the effectiveness of RTI to ensure best available use of scarce resources, increase achievement of students, and reduce the number of students requiring special education services. As a secondary goal, it is hoped that the results of this study can provide valuable information for future researchers investigating the effectiveness of Tier 2 services within the RTI framework.
Organization of the Dissertation

This introductory chapter provided the background, purpose, significance, and possible implications of this study. Chapter 2 provides, a history of special education, included as background for the development of RTI; a description of the RTI model, with a detailed outline of the various tiers; a review of research to date on the effectiveness of the RTI model, with an emphasis on Tier 2; and a history and description of the regression discontinuity technique. Chapter 3 describes the methodology used to conduct this study, and includes data sources, procedures, and data analytic techniques. Chapter 4 presents the research findings. Chapter 5 presents a discussion of the findings, the limitations, and suggestions for future studies.
The purpose of this study was to apply the regression discontinuity technique to determine the effectiveness of the RTI model of service delivery for a group of students receiving Tier 2 intervention services in the area of reading. Regression approaches were used to compare the academic performance of those students selected to receive Tier 2 intervention services in reading to those who did not receive Tier 2 intervention services. Accordingly, the literature review in this chapter contains several topics related to the research questions. First, to provide context for the RTI model and to provide background for readers unfamiliar with special education, a brief history of special education is presented. Since the development of the RTI model stemmed in part as a criticism of the discrepancy model for identifying students with learning disabilities, the IQ-discrepancy model is discussed in some detail. Third, a history and overview of the RTI model of service delivery is presented. Fourth, current research on the effectiveness of the RTI model is outlined. Finally, the regression discontinuity technique of data analysis is discussed.

*History of Special Education*

The history and evolution of education in general is replete with court cases, federal and state mandates, and social movements. For the purpose of this dissertation, a
select group of court cases, laws, and social movements, particularly relevant to the development of special education, will be presented.

The legal requirement that public school districts service students with disabilities is relatively recent, and is the product of marked efforts of parents and advocacy groups in both the legal and political domain (Yell, Rogers, & Rogers, 1998). Until the mid-1970’s, school districts were permitted by state law to refuse to enroll children who were considered ‘uneducable’ (Martin, Martin, & Terman, 1996). Those students who were enrolled in public school often received no special education services, or received services that were inappropriate for their needs. Early federal efforts to mandate services for the disabled did not initially focus on education, as the responsibility of providing education is relegated to state and local governments (U.S. Department of Education, 2008). Instead, the federal government created acts in the mid-1800’s that made grants available to states for the creation of asylums for the ‘deaf and dumb’ (Martin, Martin, & Terman, 1996). During these early years, states were relatively free to determine the manner in which disabled students were or were not, educated (U.S. Department of Education, 2008), and in fact during the 1800’s, the exclusion of disabled students from public education was widely upheld in the state courts. For example, in 1893, the Massachusetts Supreme Judicial Court determined that ‘a child who was weak in mind and could not benefit from instruction, was trouble-some to other children, and was unable to take “ordinary, decent, physical care of himself” could be expelled from public school’ (Yell, Rogers, & Rogers, 1998, p. 200). In 1919, the Wisconsin Supreme Court, in Beattie v. Board of Education ruled that a student suffering from facial contortions
which caused drooling and speech difficulties could be excluded from public school until
the fifth grade, as school officials claimed that “this condition nauseated the teachers and
other students, required too much teacher time, and negatively affected school discipline
and progress” (Yell, Rogers, & Rogers, 1998, p. 200).

Federal involvement in the modern era to improve conditions for those with
disabilities in the public schools did not resurface until the 1958, with the passage of the
National Defense Education Act of 1958, or NDEA. Spurred by the Soviet’s launching
of Sputnik, Congress passed this act, which provided among other things, grants designed
to improve the teaching of math, foreign language, and science in the elementary and
secondary grades (U.S. Department of Education, 2008). Closely following suit,
President Dwight Eisenhower proposed the passage of a small act known as Public Law
85-926, which provided financial support to institutes of higher education for training
personnel in teaching students with cognitive impairments, or mental retardation.

Key Court Cases

The civil rights movement during the 1950’s and 1960’s sought societal changes
that would extend equality of opportunity to African Americans. The civil rights
movement evolved from perhaps one of the most important judicial challenges, the 1954
court case of Brown v Board of Education (Brown Foundation). Generally, this series of
court cases maintained that the racial segregation of schools violates the 14th Amendment
of the U.S. Constitution, which states,

All persons born or naturalized in the United States, and subject to the jurisdiction
thereof, are citizens of the United States and of the State wherein they reside. No
State shall make or enforce any law which shall abridge the privileges or immunities of citizens of the United States; nor shall any State deprive any person of life, liberty, or property, without due process of law; nor deny to any person within its jurisdiction the equal protection of the laws (U.S. Constitution, 14th Amendment, Sec 1).

With the goal and attainment of ending racial segregation in schools and declaring the practice unconstitutional, the ruling on Brown v Board of Education has had far-reaching implications on the shaping of our nation’s policies regarding basic human rights, and maintained that segregation solely because of an individual’s unalterable characteristics, such as race or disability, was unconstitutional (Yell, Rogers, & Rogers, 1998).

After the judicial victory in the Brown case, parent advocates of children with disabilities began to challenge school districts, proclaiming districts ignored their constitutional obligations of equal protection of the law by failing to provide publicly supported education to disabled students (Yell, Rogers, & Rogers, 1998). In the 1972 case, Pennsylvania Association for Retarded Citizens v Pennsylvania (PARC), the plaintiffs argued that students with mental retardation were not provided with publicly supported education. PARC was resolved by agreement consenting that all children with mental retardation and between the ages of 6 and 21 must be provided with a free public education, and that it was most desirable to educate such students in a program similar to that of their non-disabled peers. A second key court case, Mills v. Board of Education (1972), found that a group of severely disabled students excluded from public education were improperly excluded without due process of law. Similar to the PARC case, the outcome of Mills v. Board of Education (1972) maintained that disabled children must be
afforded a publicly sponsored education, and put into place a variety of safeguards including: the right to a hearing with representation; the right to appeal; the right to have access to records; and the requirement of written notice throughout the process (Yell, Rogers & Rogers, 1998). All of these safeguards have since been incorporated into federal law by Congress.

The first major federal effort to subsidize direct services for selected students in public elementary and secondary school, was the Elementary and Secondary Education Act of 1965. The passage of a related law (Public Law 89-313) in the year following provided mandates that special needs children could be counted for entitlement purposes and funds earmarked for this select group of children. During the 1960’s parent advocate and special interest groups lobbied for a special administrative unit at the federal level, a bureau housed in the U.S. Office of Education. In 1966, Congress mandated the Bureau for the Education of the Handicapped (BEH), which provided grants for states to initiate, expand, and improve services for students with disabilities. Further, the BEH created a number of federal programs targeted at specific population subgroups, for example, education of children who were deaf/blind or multiply handicapped, or model programs for those children with specific learning disabilities. In the 1970’s special interest groups advocated for the adoption of an earmarking strategy that set aside fifteen percent of federal funds specifically for programs serving children with disabilities.

While the federal government was making headway in the 1960’s and early 1970’s toward providing legal provisions for servicing children with disabilities, states were relatively slower in providing such services. Although Congress passed laws
requiring states to service students with disabilities, these mandates were slow to translate into practice. Despite enacted laws, no state served all its children with disabilities, and those students who were served were often placed in inappropriate programs.

Congressional hearings found that in 1975, over one million children with disabilities were not attending school, and 3.5 million were receiving an inappropriate education (Martin, Marting, & Terman, 1996). In response, Congress passed Public Law 94-142, or the Education for All Handicapped Children Act of 1975 (EAHCA) which used federal aid to spur programming at the state and local level. Specifically, the law outlined certain procedures states must follow to receive federal funds. State departments of education were required to put in place a system of ‘child find’, to locate students with disabilities. Students suspected of disabilities were required to have evaluations conducted, in an effort to determine the nature of their disability. All children were required to have an IEP, or Individualized Education Plan in place, and services specified on the IEP were to take place in the least restrictive environment. While Public Law 94-142 was re-titled as the Individuals with Disabilities Education Act, (IDEA) in 1990, and again, as the Individuals with Disabilities Education Improvement Act (IDEiA) in 2004, the law maintains its original intent.

Of some relevance for this dissertation is the funding mechanism by which Public Law 94-142 operates. Reimbursement to school districts is based on the number of children identified with disabilities. As such, special emphasis is placed on the evaluation and identification of children with special needs, and RTI is a method schools can adopt for that purpose. While there exists agreement among politicians, researchers, and

13
diagnosticians regarding the decision criteria for certain disability categories, such as mental retardation, the decision criteria for specific learning disabilities is less clear cut and generally less agreed upon.

The general concept of *learning disabilities* has existed for the past 100 years, when it was noticed that some children with apparently intact cognitive functioning experienced difficulty reading. Since that time, the condition has been described using various labels, such as *word blindness* and *dyslexia*, until finally in 1963, when formally termed *learning disability* (LD), to which it is referred today (Aaron, Joshi, Gooden, & Bentum, 2008). In 1968, the term *learning disability* was defined and incorporated into federal law by the National Advisory Committee of the Handicapped (NACH). The definition, largely unchanged today, defines a learning disability as:

> The term “specific learning disability” refers to a disorder in one or more of the basic psychological processes involved in understanding or in using language, spoken or written, which may manifest itself in imperfect ability to listen, think, speak, read, write, spell, or do mathematical calculations. The term includes such conditions as perceptual handicaps, brain injury, minimal brain dysfunction, dyslexia, and developmental aphasia. The term does not include a learning problem which is primarily the result of visual, hearing, or motor handicaps, of mental retardation, of emotional disturbance, or of environmental, cultural, or economic disadvantage (Ofiesh, 2006, p. 884).

**IQ-Discrepancy Model**

The passage of the Education for All Handicapped Children Act in 1975 incorporated the term *learning disability* into the legislation, and officially recognized LD as a form of disability. With the passage of the act, it became necessary to develop a system of eligibility criteria by which learning disabilities could be objectively diagnosed. Because learning disabilities are typically defined as average or above
average levels of intelligence coupled with below average academic performance, it made logical sense to compare the IQ and achievement scores of children presenting academic difficulties. Inherent in this method is an assumption that, a child possesses the intrinsic ability to perform well academically, but something is holding him or her back (Harry & Klinger, 2007). If there exists a difference, or discrepancy, between a child’s level of cognitive ability and level of academic achievement, a learning disability may account for this difference. As with the current changes to IDEA 2004, which are designed to help operationalize the definition of learning disabilities, a separate set of regulations in 1977, known as the Federal Register, was developed. These regulations state that a student has a diagnosis of SLD if (a) the student does not achieve at the proper age and ability levels in one or more specific areas when provided with appropriate learning experiences, and (b) the student has a severe discrepancy between achievement and intellectual ability in one or more these seven areas: oral expression, listening comprehension, written expression, basic reading skills, reading comprehension, mathematics calculation, and mathematics reasoning (Ofiesh, 2006).

This method by which learning disabilities have been defined and diagnosed has become known as the **IQ-Discrepancy Model** for determining learning disabilities (Graner, Faggella-Ludy, & Fritschmann, 2005). In determining eligibility for services as a student with a learning disability, a primary requirement was that students exhibit a ‘severe discrepancy’ between ability and achievement not primarily the result of visual hearing or motor disabilities; low cognitive functioning; emotional disturbance; or environmental, cultural, or economic disadvantage. One of the many criticisms of this
method for identification and eligibility is that ‘severe discrepancy’ is poorly operationalized and implemented across states, districts, schools, and individual students (Scruggs and Mastropieri, 2002). A widely adopted criterion used by many school districts is the ‘two standard deviation rule’ (i.e., if a student exhibited a two-standard deviation difference between ability and achievement in at least one academic area, the student could be determined to have a learning disability). The difficulty with using this method is that students characterized as having low average levels of cognitive ability are almost guaranteed to fail to meet this requirement. In the past, these students have been considered ‘performing at their level of ability’ and were often denied services.

**History of RTI**

During the 1990s, researchers began to notice and address problems in the general and special education systems. Areas of concern included the discrepancy between general and special education delivery, lack of emphasis on early intervention, little emphasis on research-based practice and techniques, and especially, the limitations of the traditional IQ-achievement discrepancy model for identifying children with learning disabilities (Martinez, Nellis, & Prendergast, 2006). At that time, the IQ-achievement discrepancy model was conceptualized as the existence of a ‘severe discrepancy’ between a student’s level of cognitive ability and performance on a standardized assessment of academic achievement (Graner, Faggella-Ludy, & Fritschmann, 2005). In 2001, the U.S. Department of Education Office of Special Education Programs sponsored a summit to discuss in particular, the use of the IQ-achievement discrepancy model and limitations in
identifying students with learning disabilities. A paper among the series produced included an article entitled *Responsiveness to Intervention: An Alternative Approach to the Identification of Learning Disabilities* by F. Gresham (2001). Gresham’s work was critical, as it introduced the concept of RTI into discussion on education reform. One of Gresham’s key points involves the application of the LD label: identifying a child with a learning disability should occur only after the child’s failure to respond to a scientifically-validated intervention. Following the summit, President George W. Bush called upon the Commission on Excellence in Special Education to propose new policies for special education law. The new policies highlighted three points relevant to the introduction of RTI: (a) the importance of educational outcomes over process, (b) the prevention of academic and behavioral difficulties, and (c) the consideration of students with disabilities as general education students first (Martinez et al., 2006). In 2004 these points were incorporated into the reauthorized Individuals with Disabilities Education Act of 1997 as the Individuals with Disabilities Education Improvement Act of 2004. According to these new regulations, a State must adopt, consistent with 34 CFR 300.309, criteria for determining whether a child has a specific learning disability as defined in 34 CFR 300.8(c)(10). In addition, the criteria adopted by the State:

- Must not require the use of a severe discrepancy between intellectual ability and achievement for determining whether a child has a specific learning disability, as defined in 34 CFR 300.8(c)(10);
- Must permit the use of a process based on the child’s response to scientific, research-based intervention; and
• May permit the use of other alternative research-based procedures for determining whether a child has a specific learning disability, as defined in 34 CFR 300.8(c)(10).

In essence, these final regulations state that in identifying the presence of a specific learning disability, states must not mandate the use of the IQ-achievement discrepancy formula, but must also permit the use of a process that assesses the child’s response to research-based intervention techniques, now known as RTI (Martinez et al., 2006).

**Overview of RTI**

There is not currently a formal definition for RTI, nor is there a singular model that is established or widely endorsed by researchers and educators (Fuchs, Fuchs, Compton, Bouton, Caffrey, & Hill, 2007; Cortiella, 2008; Barnes & Harlacher, 2008). Generally, RTI is conceptualized as a model of service delivery that, using a problem solving approach, seeks to address student difficulties using efficient research-based instruction through increasingly intensive levels of support. While some variation exists in the conceptual model of RTI service delivery, most researchers and educators agree with the overarching philosophy, as described here. First and foremost, the overarching philosophy of RTI focuses on preventing academic and behavioral problems through early identification, remediation, and use of research-based educational practices, and secondly, on identifying students with disabilities (Fuchs, Fuchs, Compton, Bouton, Caffrey, & Hill, 2007). Thus, RTI serves two purposes: (a) to immediately address and
remediate learning difficulties, and (b) as an alternate method from the IQ-discrepancy model for diagnosing learning disabilities.

Within the scope of prevention, RTI consists of a school-wide service of delivery that seeks to remediate academic and behavioral difficulties before they escalate into such severity that remediation is difficult or costly. This service of intervention is available to all students, and may include professional development, enhanced classroom instruction, supplemental instruction, or school-wide behavioral supports (Denton, Fletcher, Anthony, & Francis, 2006). With prevention methods in place, those children who continue to struggle are identified as needing intensive supports, which may include the provision of special education services. Thus, RTI may be expanded to allow for the differentiation between two explanations for low academic achievement: poor instruction or the presence of a disability (Fuchs, Fuchs, & Hollenbeck, 2007).

Graner, Faggella-Luby, Nanette, and Fritschmann (2005) described an RTI “archetype,” composed of eight primary features (high-quality classroom instruction; research-based instruction; classroom performance measures; universal screening; continuous progress monitoring; research-based interventions; progress monitoring during interventions; and fidelity measures) and five secondary features (multiple tiers; transitioning from least-intensive to most-intensive instruction; implementation of differentiated curriculum; instruction delivered by paraprofessionals and other staff; varied duration, time, and frequency of interventions; and categorical or non-categorical placement decisions). This model of service delivery generally is conceptualized on a continuum that ranges in intensity in a tiered or phase format, although variations exist in
both the number and nature of tiers utilized and is one of the criticisms of RTI (Mastropieri & Scruggs, 2005; Barnes & Harlacher, 2008). Typically, RTI is conceptualized as a three-tiered model in which “instruction is layered over time in response to students’ increasing needs” (Vaughn, 2003), but four-tiered models have been utilized in some districts across the nation (Tilly, 2003). The three-tiered model is relatively well documented in the literature (Martinez et al., 2006; Vaughn & Roberts, 2007; Stecker, 2007) and is the model most relevant for this dissertation.

The 3-Tiered Model

**Tier 1.** Tier 1, sometimes identified in the literature as the primary level of service delivery, includes those universal interventions and programs that are accessible to all students and support positive academic and behavioral outcomes (see Figure 2.1). Generally, Tier 1 is considered the general, or regular academic curriculum. At the heart of Tier 1 is the incorporation of core curricula and instruction backed by well-designed, scientifically-based research methods, and demonstrated as effective in increasing educational outcomes. For example, in the area of reading, the core program often contains research-based practices in the areas of phonemic awareness, alphabetic understanding, fluency, vocabulary, and comprehension (Marston, 2005; Vaughn, 2003). Within Tier 1, grouping is flexible and classroom-centered, and students receive 90 minutes or more of core academic instruction per day (Marston, 2005). With a sound, research-based curriculum in place, school personnel must then gather data on the school-wide effectiveness of the programs as they operate within the general education setting.
This data-gathering procedure is critical to the implementation of the RTI model, and often takes the form of large-scale screening of all students (Barnes & Harlacher, 2008). Typically, students are assessed two or three times per year via benchmark assessments in core academic areas to gather critical data about all students (Vaughn, 2003; Tilly, 2003; Barnes & Harlacher, 2008). *Benchmark assessments* are “reliable, valid, standards-based assessments administered to a whole group of individuals at regular intervals….the results can be used to determine student growth and student performance relative to statewide grade-level achievement expectations” (Pennsylvania Training and Technical Assistance Network [PaTTAN], 2007). Within Tier 1, the benchmark assessments serve a dual purpose: assisting in guiding classroom instruction and identifying students

![Figure 2.1 Response to Intervention Model of Service Delivery (Nebraska Department of Education, 2008).](image-url)
in need of intervention or re-teaching. Students who are not making satisfactory progress according to pre-established criteria and are determined “at-risk” can be identified and targeted for more intensive intervention. Interventions administered in Tier 1 are mostly “large scale” and commonly are administered by the general education teacher and in the general education setting. In summary, Tier 1 interventions consist of data-driven practices utilized by the regular classroom teacher to identify children who are considered at-risk and in need of more intensive intervention (Hilton, 2007; Vaughn, 2003).

 Tier 2. Tier 2, (see Figure 2.1) often referred to in the literature as the secondary level of service delivery, targets the identified “at-risk” students who are struggling with the core academic curriculum (Vaugn & Roberts, 2007). This tier marks the turning point: it is meant to be the primary vehicle for selecting regular education students in need of intervention, providing the intervention, and transitioning those students back into the regular or general education setting (Fuchs, Compton, Fuchs, Bryant, & Davis 2008). The specific needs of each student, or groups of students, are identified through deficiencies on the benchmark assessments, and research-based interventions are implemented. As described by Martinez et al. (2006) these interventions are applied through one of two methods: a standard-protocol approach, where all children whose data revealed difficulty in similar areas are administered the same empirically-validated intervention, or a problem-solving model, which involves the use of building-level teams to review data and select appropriate intervention strategies. Once an intervention has been identified, at-risk children receive small-group or individual supplemental instruction (over and above the 90 minutes of core instruction provided in
the classroom setting), typically for a minimum of 30 minutes per day. These interventions are designed to not only supplement, but also to support instruction received in the general education setting. Secondary interventions provided in Tier 2 may be implemented by teachers, trained paraprofessionals, reading specialists, or tutors, provided they receive ongoing professional development and support (Vaughn & Roberts, 2007). The setting for the intervention may occur within or outside the classroom setting, as deemed appropriate. The importance of progress monitoring becomes readily apparent in Tier 2; students are monitored carefully and systematically over the course of the intervention period – anywhere from weekly, in most cases, to twice monthly. The length of Tier 2 instruction is being researched continually; some studies suggest that 10 to 12 weeks or 50 sessions of instruction are sufficient to make a decision about the student’s progress (Vaughn & Roberts, 2007). Students making adequate progress at the end of an intervention period may return to the general education curriculum; students not making acceptable progress may receive an amended intervention at the secondary level, or may proceed to the next tier. Vaughn (2003) recommends that students proceed to the next tier if any of the following apply: (a) the student has participated in two rounds of Tier 2 instruction and has not made sufficient progress after the intervention has been modified, for a total of 20 weeks or 100 sessions; (b) if the student has demonstrated a marked lack of progress after one round of Tier 2 and further instruction in Tier 2 is considered insufficient; or (c) the student has received previous Tier 3 instruction and has returned to Tier 2.
Tier 3. Tier 3 (see Figure 2.1), often referred to in the literature as the tertiary level, is more intensive and individualized than Tier 2, and students who continue to struggle with the additional support provided in Tier 2 are referred for this third level of service (Stecker, 2007). Intervention in Tier 3 is markedly more individualized and supplemental, with progress monitoring more frequent. Homogeneous, small-group, or individualized instruction is administered to these students, for a minimum of two 30-minute sessions per day in addition to the 90 minutes of core instruction provided in Tier 1. Progress monitoring is conducted at least twice per month, but more often weekly, and in some cases, daily. The duration of the intervention in Tier 3 is considerably longer, and may span months or perhaps, years. Students who fail to make adequate progress with intensive intervention are often referred for special education evaluation to rule out other disabilities, such as a cognitive disability or emotional disturbance (Fuchs, & Fuchs, 2005). The transition from intervention services spanning these three tiers and provided in the “regular” or “general” education setting to accommodations and modifications provided in a special education setting can occur in Tier 3 or in Tier 4, if a four-tiered model is conceptualized (Mastropieri & Scruggs, 2005).

Potential Benefits of RTI

The proponents of RTI cite many potential benefits of the model for addressing learning difficulties. As outlined by Compton, Fuchs, Fuchs, and Bryant (2006) the big promises of RTI can be summarized as follows: (a) an earlier identification of learning disabilities to avoid a “wait to fail” model, (b) a strong focus on providing effective instruction and
improving student outcomes, and (c) a decision-making process supported by continuous progress monitoring of skills closely aligned with desired instructional outcomes.

The notion of ‘wait-to-fail’ is reflected in the traditional, IQ-discrepancy model whereby students who are not performing ‘poorly enough’ on a standardized assessment to qualify are delayed special education services. An illustration of ‘wait-to-fail’ is provided in a hypothetical example: Take, for example, Student ‘A’, a student with average to above average IQ, and below average academic functioning in basic reading. Under the IQ-discrepancy model, this particular student would not be performing at his or her level of potential; that is, a student possessing average to above average IQ should, in theory, possess the intelligence to perform academically at that same level. Provided other reasons for poor performance, such as lack of instruction or cultural disadvantage are eliminated, this student under the IQ-discrepancy model would qualify for special education services as a student with a Specific Learning Disability. Now consider another student, Student ‘B’, with below average IQ and performing at the same level of achievement as Student ‘A’ in basic reading. Both students are performing at the same level of achievement, but differ in ability. Under the IQ-Discrepancy model, Student ‘B’ is not performing ‘poorly enough’; that is, Student ‘B’’s academic performance, although low, is commensurate with level of cognitive ability. There exists no severe discrepancy between ability and achievement, and the diagnosis of a Specific Learning Disability is delayed, perhaps for a year or so, until the student falls behind even further. The RTI model eliminates the ‘wait-to-fail’ approach, as students’ needs are addressed immediately.
The second and third ‘big promises’ of RTI, effective instruction for all students and use of an objective, data-driven decision making process, are intertwined and reflected in the tier approach. The incorporation of scientifically-based instructional methods at each tier leads to high-quality instruction for every student. The incorporation of benchmark assessments, administered to all children, guarantees that the progress of each student is continually monitored. Frequent assessment is critical, as current educational philosophy rests on the premise that continuous assessment fosters skill improvement (Brown-Chidsey & Steege, 2005). The decision to advance students through tiers is based not on opinion or vague criteria, but using data-driven practices, of which frequent assessment is a critical piece.

Criticisms of RTI

While RTI has many proponents, several researchers believe that RTI is appropriate as only a method for providing intervention. These researchers cite the importance of differentiating between RTI as a method for addressing low student achievement, and RTI as a method for identifying or diagnosing learning disabilities (Hale, Kauffman, Naglieri, & Kavale, 2006; Ofiesh, 2006). IDEIA 2004 permits RTI processes as a potential method to not only support student learning, but also for diagnostic purposes, a practice viewed by some researchers as problematic. Ofiesh (2006) writes of the potential danger of taking comprehensive evaluations (e.g., standardized assessments, such as IQ tests) out of the process. In the article, the author asks a compelling question:
Is it our responsibility as educators to identify a student as one who has a learning disability (i.e., as opposed to, e.g., a developmental delayed or “other health impairment”) or simply to provide all children an adequate opportunity to learn? If the answer is “yes,” that we should identify a student’s learning disability as well as provide an opportunity to learn, then we must be cautious about RTI models that do not provide a measure of psychological or cognitive functioning at one of the tiers, as some have proposed. To identify an SLD as defined by the law and by those who originally defined the construct of SLD, we must be able to understand the underlying cause of instructional problems, not just an individual’s response to instruction (p. 883-884).

Other researchers consider the RTI process as somewhat vague. McBride, Dumont, and Willis (2004) provide a litany of challenges to proponents of RTI to (a) provide methods for determining whether interventions are empirically based; (b) establish criteria for determining whether children are in need of RTI; (c) formalizing the RTI process; and (d) determining the duration of RTI before classifying children as learning disabled.

Research on Effectiveness of Tier 2

Much of the research on RTI reflects the differentiation of RTI as a means for addressing lack of achievement versus a method for diagnosing learning disabilities. To date, research studies have taken one of two approaches: contrasts of student performance as students receive interventions and move through the various tiers, and outcome studies regarding the number of students later identified as requiring special education services (Marston, 2008). Many studies of student performance use post-program, norm-referenced achievement tests or pre- and post-program slope comparisons to address intervention effectiveness, and there exists much research searching for validated
interventions (Burns, Appleton, and Stehouwer, 2005). Studies of specific interventions are beyond the scope of this dissertation and will not be presented here.

A recently published study investigating the effects of an RTI approach to screening and eligibility determination was conducted by VanDerHeyden, Witt, and Gilbertson (2007). In this study, the researchers found that use of RTI resulted in fewer special education evaluations (29 children prior to RTI implementation and 14 children after RTI implementation). Further, more of those students who were evaluated for special education services were determined eligible for those services. Following the first year of implementing RTI practices, specific learning disability diagnosis decreased from 6% of elementary school children to 3.5% of elementary school children, district-wide.

A study conducted by Bryant, Bryant, Gersten, Scammacca, and Chavez (2008) investigating the effects of Tier 2 intervention on the performance on first- and second-grade students in math, revealed mixed results. In this study, 26 first-graders and 25 second-graders received tutoring sessions in same-ability, small groups consisting of three to four students. Tier 2 instruction consisted of 15-minute tutoring sessions, administered three to four days per week and targeting basic mathematic skills. Length of Tier 2 instruction spanned 18 weeks, with a median of 64 sessions for first graders and 62 sessions for second graders. Using regression discontinuity to analyze the data, the researchers found differential effects by grade level. For Tier 2 first-grade students, the regression discontinuity did not reveal a program effect, whereas for Tier 2 second-grade students, the analysis demonstrated a significant program effect. The researchers
suggested that ‘additional intervention time may be needed to allow students more practice opportunities with different representations’ (p. 30), and suggest that the length and breadth of intervention sessions needed to be increased in Tier 2 instruction.

In effort to determine the effectiveness of Tier 2 instruction in reading, Fuchs, Fuchs, Compton, Bryant, and Davis (2007) conducted a longitudinal study of first-grade students receiving secondary intervention in sight word recognition and letter-sound recognition. Tier 2 instruction was defined as small group tutoring, four sessions per week, for 45 minutes per session. In this study, 252 children, considered the poorest readers in the cohort were selected to participate in the study. Children were randomly assigned to one of three groups: (a) fall tutoring (n = 84), where all students participated in small group tutoring for 9 weeks during the fall semester; (b) spring tutoring (n = 84), where all students considered non-responsive to the fall tutoring participated in small group tutoring for 9 weeks during the spring semester; and (c) control students (n = 84), who received no tutoring but were matched to non-responders in the spring group. All groups were administered weekly progress monitoring probes, consisting of fluency measures of word identification, across the 18 weeks of the study. Intercept and slope terms for the tutored and control groups were calculated. Slope 1, corresponding to growth prior to the spring tutoring, and slope 2, corresponding to growth after spring tutoring, were compared. To explore whether changes in slope 2 were a result of tutoring, ‘group’ (dummy coded) was used to predict slope 1 and slope 2. Findings reveal that ‘group’ was not a significant predictor for the intercept or slope 1, suggesting that growth for the tutored and control groups were similar during the fall semester. In contrast,
‘group’ was a significant predictor for slope 2, suggesting that word identification fluency was greater in the tutored group that the control group.

Another study investigating the effects of a tiered format to addressing reading problems was conducted by O’Connor, Harty, and Fulmer (2005). In this study, a selected group of at-risk Kindergarten students, selected to receive Tier 2 and Tier 3 services, was tracked from Kindergarten to third grade. In the second month of Kindergarten, students were administered measures of phoneme awareness and letter knowledge. Students were considered ‘at-risk’ if they named fewer than 15 letters or 10 phonemes. The ‘at-risk’ students (n = 31) were selected to receive Tier 2 services, consisting of small-group supplemental instruction, administered 10 to 15 minutes, three times per week. At the end of Kindergarten, 15 of the initial ‘at-risk’ students were performing in the ‘average’ range, but 6 of them needed additional support in first or second grade. Nine of the initial ‘at-risk’ children in Kindergarten remained in the ‘average’ range over the course of the next three years on both word-level skills and fluency. In first grade, 16 of the original 31 ‘at-risk’ students were again selected to receive Tier 2 intervention. For first grade, Tier 2 services were delivered as small-group instruction, for 20 to 25 minutes, three times per week. By December of first grade, two students performed in the ‘average’ range and maintained that level through third grade. The remaining students transitioned back and forth between Tier 2 and Tier 3 services over the next two years; nearly two-thirds of the initially identified ‘at-risk’ group received either Tier 2 or Tier 3 instruction from Kindergarten through third grade. The authors cite the ‘feasibility’ of implementing Tier 2, as:
Tier 2 was delivered only three days per week for 15 to 25 minutes per session…11% of our sample needed more intensive instruction (5 days per week for 30 minutes, sometimes delivered one-on-one) for twenty weeks to three years, and even with Tier 3, 8% were identified for special services by the end of third grade (O’Connor et. al., 2005, p. 537).

**Relevant Research Using Regression Discontinuity**

The regression discontinuity technique is generally under-utilized in education (Gersten & Dimino, 2006) and especially within the context of RTI. To date, a search on ERIC produced a single research study that utilized regression discontinuity to investigate the effectiveness of RTI. The study, conducted by Bryant, Bryant, Gersten, Scammacca, and Chavez, (2008) investigated the effects of a Tier 2 math intervention for first- and second-graders using the regression discontinuity technique. In this study, 100 first-grade students who did not qualify for Tier 2 math intervention and 26 first-grade students who did qualify for Tier 2 math intervention were administered a pre- and post-program measure of mathematics achievement. Similarly, 115 second-grade students who did not qualify for Tier 2 services and 25 second-grade students who did qualify for Tier 2 services were administered a pre- and post-program measure of mathematics achievement. The researchers found differential results of effectiveness by grade level; for first-grade Tier 2 students, the analysis revealed no effect of program \( (b = .04) \), whereas for second-grade Tier 2 students, the analysis revealed a positive program effect \( (b = .19) \).

Outside of the context of RTI, regression discontinuity has been utilized to investigate the effects of various educational variables, such as educational training (Malamud, & Pop-Eleches, 2008), retention (Jacob, Brian, & Lefgren, 2007), remediation
(Calcagno, Carlos, Long, & Bridget-Terry, 2008; Leake & Lesik, 2007; Jacob, & Lefgren, 2001), and teacher variables (Jacob & Lefgren, 2004). Regression discontinuity has also been utilized within the context of educational program evaluation (Gormley, Gayer, Phillips, & Dawson, 2005; Wong, Cook, Barnett, & Jung, 2008). These topics are beyond the scope of this dissertation, and will not be described here.

Chapter 2 provided: a history of special education, included as background for the development of RTI; a description of the RTI model, with a detailed outline of the various tiers; a review of research to date on the effectiveness of the RTI model, with an emphasis on Tier 2; and a history and description of the regression discontinuity technique. Chapter 3 will address the research methodology used to address the topics of interest. Description of participants, selection criteria, measures, and method of data analysis will be presented in Chapter 3.
Chapter 3
Method

The purpose of this study was to apply the regression discontinuity technique to determine the effectiveness of the RTI model of service delivery for a group of first-grade students receiving Tier 2 intervention services in reading. A second purpose was to determine the appropriateness of the regression discontinuity technique in studying the RTI model. This chapter begins by providing a discussion of the source of data, measures used to track progress, and the intervention program used by a school district. A detailed discussion of the research design, along with the statistical model, is also presented.

Source of Data

Data used for this study consisted of archival data collected during the 2007-2008 school year by a large urban district located in southwestern Ohio. During the 2007-2008 school year, the district collected benchmark data on students as part of its newly-adopted RTI process. The data file contained 141 pairs of first-grade students’ Oral Reading Fluency (ORF) subtest scores on the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) collected during the winter (January) and spring (March) benchmark assessments and a group membership identification code (1 if Tier 2, 0 if control). ORF
scores are numerical values representing the number of words read correctly within a one-minute period. No identifying or demographic information was included in the data file. However, information contained on the Ohio Department of Education website (www.ode.oh.gov) states that the average daily enrollment for the district during the 2007-2008 school year was 3450 students. District-wide, students were characterized as predominantly Caucasian (82.2%) and economically disadvantaged (52.5%). Race/ethnicity and economic information for the students participating in this study was not included in the data file and although requested, was not provided by the district. Student gender was provided by the district school psychologist and was reported in aggregated form. Subjects for this study included 141 first-grade students; 51% of students (n = 72) are females, and 49% (n = 69) are males. Quantitative and qualitative data (i.e., information regarding data collection, method of assignment, type of intervention used, and length of intervention period) were provided to the researcher by the district’s school psychologist, who obtained information from teachers on the researcher’s behalf. Permission to receive and use data, and permission to speak with the district’s school psychologist was granted by the district’s curriculum coordinator and the school psychologist; however, permission to directly contact teachers for information was requested by the researcher, but denied by the school district.

*Measures for Pre- and Post-Assessment*

Within the RTI framework, benchmark assessments are used as a means of identifying at-risk students and tracking their progress. A series of reading assessments
commonly used for benchmark purposes and developed in recent years by the University of Oregon Center on Teaching and Learning is the Dynamic Indicators of Basic Early Literacy Skills, also known as DIBELS (http://www.dibels.uoregon.edu). DIBELS is founded on curriculum based measures (CBM), developed by Deno in the 1970’s and 1980’s (Deno, 1985), and is best described by that author (Deno, 1992):

The purpose of CBM is to enable teachers to improve student performance. The primary assumption is that CBM will be used to create a data base for each student to allow the teacher to evaluate the effectiveness of an individual student's educational program. The goal of this individual student monitoring is to create a formative evaluation framework within which the teacher can systematically test alternative approaches to instruction for individual students. Our hypothesis has been that if teachers create and use a data base for empirically testing alternative approaches to instruction, they ultimately will be more effective in helping students attain proficiency in basic skills (p. 5).

The DIBELS is designed to be a quick and reliable system for measuring reading fluency skills, and includes fluency measures of reading components identified by reading specialists as critical for reading development (Schilling, Carlisle, Scott, & Zeng, 2007). Referred to on the DIBELS website as the ‘5 Big Ideas in Basic Reading’, this set of reading components consists of phonemic awareness, alphabetic principle, accuracy and fluency with connected text, vocabulary, and comprehension (University of Oregon Center on Teaching and Learning). Each subtest incorporated in the DIBELS reflects one of the core components of reading. All subtests are a standardized, individually-administered curriculum-based measure consisting of short, one-minute administrations of a core reading component, which vary by grade level (University of Oregon Center on Teaching and Learning). For this study, parallel forms of the Oral Reading Fluency
(ORF) subtest of the DIBELS were administered during the winter and spring assessment.

Generally, oral reading fluency refers to the ability to read connected text smoothly, effortlessly, and with little focus on the mechanics of reading, such as decoding (Mather & Goldstein, 2001). Research has demonstrated that oral reading fluency has strong predictive validity for overall reading skill. In a study investigating the predictive validity of the Oral Reading Fluency subtest and the Iowa Test of Basic Skills, first grade winter and spring correlations were .69 and .75, respectively, with the addition of the Oral Reading Fluency subtest accounting for the largest percentage of variance (Schilling, Carlisle, Scott, & Zeng, 2007). In an investigation of the predictive validity of first grade oral reading fluency scores to a measure of overall reading in second grade, Stage (2001) reported a coefficient of .849, indicating that oral reading fluency scores in first grade significantly predicted second grade reading scores. Other research demonstrates that oral reading fluency has predictive validity for future difficulties in reading comprehension on state reading assessments (Shapiro, Solari, & Petscher, 2008) and on measures such as the Metropolitan Achievement Tests (Hixson & McGlincey, 2004).

The DIBLES ORF subtest consists of a one-minute assessment of reading fluency, defined as the number of words read correctly within a one-minute time period. Substituted or omitted words are counted as errors, as are hesitations of more than three seconds. Words self-corrected within the three-second criteria are scored as accurate. The
number of words read aloud minus the number of errors represents the Oral Reading Fluency score (University of Oregon Center on Teaching and Learning).

Reliability and validity of the ORF has been demonstrated in various studies. Test-retest reliabilities for elementary students ranged from .92 to .97; alternate form reliability of different reading passages drawn from the same level ranged from .89 to .94 (Tindal, Marston, & Deno, 1983). Criterion-related validity studied in eight separate studies in the 1980's reported coefficients ranging from .52 to .91 (Good & Jefferson, 1998).

Administration procedures are detailed on the DIBELS website (https://dibels.uoregon.edu/measures.orf.php). A sample ORF probe contains the following directions:

When I say “begin”, start reading aloud from the top of the page (point).

Read across the page (point). Try to read each word. If you come to a word you don’t know, I’ll tell it to you. Be sure to do your best reading. Ready, begin. At the end of one minute, place a bracket (]) after the last word, and say, “Stop”.

The examiner starts the stopwatch when the student says the first word of the passage. If the student cannot read the first word within the three-second time-limit, the examiner tells the student the word, and counts the word as incorrect. Words read incorrectly are marked with a slash. At the end of the one-minute period, a bracket (]) is placed after the last word read by the student, and the passage scored for numbers of items read correctly.
Assignment of Participants

The ORF subtest of the DIBELS was administered to all first-grade students in the middle of the 2007-2008 school year and at the end of the 2007-2008 school year. The middle-of-year administration will be referred to as the ‘winter’ benchmark, and the end-of-year administration will be referred to as the ‘spring’ benchmark. Special education and former Title I teachers administered the ORF subtest. Prior to administering the winter and spring ORF measures, teachers were trained on administration and scoring procedures by school psychologists and educational staff housed at a local Educational Service Center. Students considered at-risk according to their performance on the benchmark and percentile rank relative to their peers were reviewed by the school’s RTI team. The RTI team was comprised of the student’s classroom teacher, building principal, district school psychologist, parent, and ‘core’ RTI team members (i.e., teachers who attended all meetings, even though they may not have been assigned that particular student). Students were selected to receive supplemental assistance in reading, or Tier 2 intervention, according to at-risk status on the winter benchmark ORF assessment and their relative percentile ranking; students scoring at or below the 10th percentile and who were considered to fall in the ‘at-risk’ or lower end of the ‘some risk’ category were selected to receive Tier 2 services. All students were monitored over the course of three months, after which time first-grade students in both the intervention and non-intervention group were administered the spring benchmark assessment, a parallel form of Oral Reading Fluency measure of the DIBELS.
DIBELS assigns each student into a ‘category of risk’ according to their established benchmark criteria. Categories of risk on the DIBELS benchmark assessments were formulated by:

following a large group of students in a longitudinal manner to see where students who were "readers" in later grades were performing on these critical early literacy skills when they were in Kindergarten and First grade so that we can make predictions about which students are progressing adequately and which students may need additional instructional support (https://dibels.uoregon.edu/faq.php).

The DIBEL ORF scores are then classified into one of three categories of risk: at-risk, some risk, and low risk, and describe the probability that a student will fail to meet subsequent reading benchmarks. The creation of risk categories is described on the DIBELS website:

Benchmark goals for each measure and time period were established using a minimum cut point at which the odds were in favor of a student achieving the next benchmark goal. For a score to be considered a benchmark goal, at least 80% to 85% of students in the sample with that score at that point in time had to achieve the next goal. So, for a child with a score at or above the benchmark goal at a given point, the probability is high for achieving the next goal; the probability of need for additional support to achieve the next goal is low (https://dibels.uoregon.edu/dibelsinfo.php).

Table 3.1 provides an overview of the three classifications and the corresponding range of scores assigned for each category of risk (for example, a student with an ORF score of 5 on the winter administration would fall within the ‘at-risk’ category, and be considered at a relatively high risk for failure to achieve benchmark at the next administration; a student with an ORF score of 45 on the spring administration would be considered ‘low-risk’ for failing to meet benchmark expectations).
Table 3.1 Benchmark Scores Corresponding to Classification of Risk on the DIBELS

<table>
<thead>
<tr>
<th>Category</th>
<th>Winter Score</th>
<th>Spring Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>At Risk</td>
<td>0 - 7</td>
<td>0 - 19</td>
</tr>
<tr>
<td>Some Risk</td>
<td>8 - 19</td>
<td>20 - 39</td>
</tr>
<tr>
<td>Low Risk</td>
<td>20 and above</td>
<td>40 and above</td>
</tr>
</tbody>
</table>

After reviewing each student’s ORF scores on the winter administration, the classroom performance of lower-achieving students was discussed during grade level team meetings. Struggling students were selected to receive Tier 2 services based on the following criteria: the winter benchmark score on the ORF subtests fell in the ‘at-risk’ or lower end of the ‘some risk’ category, according to criteria on the DIBELS; the student’s performance on the ORF subtest fell below the 10th percentile, when compared to the entire group of first grade students; and the student’s teacher felt that Tier 2 services were necessary. Out of 141 first-grade students, 21 met criteria for Tier 2 services. Table 3.2 provides a breakdown of the students in each group, by gender. The intervention group contains more boys (57%) than girls (43%). Conversely, the control, or

Table 3.2 Student Group by Gender

<table>
<thead>
<tr>
<th>Group</th>
<th>Tier 2</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 21</td>
<td>n = 120</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>12</td>
<td>57</td>
</tr>
<tr>
<td>Female</td>
<td>9</td>
<td>63</td>
</tr>
</tbody>
</table>
non-intervention group, contains more girls (53%) than boys (47%).

**Tier 2 Intervention Program**

Two intervention specialists, both certified special education teachers, administered the Tier 2 intervention. The intervention procedure followed a standard protocol procedure. In a standard protocol procedure, “all children whose data indicate difficulty in a certain academic area are given the same intervention that has been empirically validated to promote progress in that academic area” (Martinez et al., 2006, p 4). The interventionists incorporated a published reading series, *Reading Intervention for Early Success* (referred to in this study as *Early Success*), and a researched-based intervention procedure, known as ‘word boxes’, in the intervention program. Students participated in small-group (no more than six children per group), daily supplemental instruction. Since the district has incorporated Tier 2 into the building schedule, the length and duration of Tier 2 instruction was standardized among teachers so that students received 30 minutes of supplemental instruction per day. In the spring of the 2007-2008 school year, students were administered the post-program measure, a parallel form of the ORF subtest on the DIBELS. Out of 141 students assessed, 18 scored in the ‘at-risk’ range; 45 scored in the ‘some risk’ range; and 78 scored in the ‘low risk’ range. For ease of interpretation, refer to Table 3.2 for the lists of benchmark scores by risk for first-grade students on the winter and spring administration.
Houghton Mifflin’s *Reading Intervention for Early Success* (Early Success) is a research-based reading intervention program for struggling readers in first and second grades (Taylor, 2009). The program is designed to serve small groups of children (5 to 7 per group) for 30 minutes of supplemental instructional time. Lessons contained with the program comprise three main tasks: rereading for fluency; reading the books of the week, and working with words/writing sentences. The primary goal of the program is to ‘help students who are at-risk of reading failure succeed in reading in first grade’ (Taylor, 2009, p. 3). The reading program contains repeated readings of and guided writing about short picture books, and also contains phonemic awareness training and coaching in word-recognition strategies (Taylor, 2009). Materials include 36 picture books, summaries of the pictures books on a chart and in booklet form, and 14 easy-to-read picture books. Books range from 40 to 200 words in length and are divided into levels according to length. The program is designed so that students engage in a regular, structured, daily routine. Description of the program procedure is described in Taylor, Short, Frye, and Shearer, 1992:

Children spend three days on each story summary. On the first day, the teacher reads the original picture book to the entire class. Modeling fluent reading and appropriate book-reading behavior, she then reads the summary of the story from a chart…through teacher questioning and practice, children learn to use context as well as phonic and syntactic clues to identify words during the reading of their EIR stories. In addition, the development of students’ phonemic awareness is a particularly important part of the program. The teacher stops at four of five appropriate words to model how to segment the words into phonemes and blend the phonemes together. Students further develop their phonemic segmentation and blending ability and phonics knowledge on day one by writing up to five words
from the story with the teacher’s modeling and guidance. Children write the phonemes from the word in a series of boxes duplicated on a sheet of paper, placing one sound per box (p. 594-595).

Previous studies found that the Early Success program increases reading achievement in at-risk first grade students. In one study, sixty-seven percent of the lowest performing (below the 20th percentile) children in reading were reading on at least a pre-primer level by the end of first grade. Further, fifty percent of these children were reading at an end-of-first-grade-level or better (Taylor, 1992).

**Word Boxes**

A wealth of research supports that a common cause of reading difficulties among children is the inability to make associations between a letter and its corresponding sound (Ehri, 2001; Joseph, 2002; Phillips, Clancy-Menchetti, & Lonigan, 2008). Word boxes focus on phoneme isolation, which requires recognizing individual sounds in words (Ehri, 2001). A meta analysis of studies investigating the relation of phonemic awareness to reading difficulties revealed overwhelmingly that phonemic awareness is one of the best predictors of how well children learn to read. In a study of investigating the predictive validity of phonemic awareness and other pre-reading skills of Kindergarteners, it was found that phonemic awareness was the top predictor, correlating $r = .66$ with reading achievement in Kindergarten and $r = .62$ with reading achievement in first grade (Ehri, 2001). The use of word boxes is a phonics-based approach that “involves teaching phonemic awareness, making letter-sound associations, and teaching spelling through the use of well-established behavioral principles” (Joseph, 2002, p. 122).
Research demonstrates that word boxes are effective in facilitating and maintaining performance of first grade students in spelling and reading consonant-vowel-consonant words (Joseph, 2002) and in performing word identification and spelling tasks (Joseph, 2000).

For this study, students focused on isolating phonemes, or the smallest units that comprise spoken language and are combined to create syllables and words. Sound boxes were created by having students draw a large rectangle on a piece of blank paper. Students were then instructed to divide the rectangle into equal parts, using vertical lines to separate the squares, or boxes. The number of boxes needed depended upon the number of phonemes that could be heard in the word; for example, the word ‘cat’ has three phonemes, and would require three boxes. Next, the teacher presented the word to the students. The teacher modeled how to stretch the word, saying it slowly and emphasizing the sounds of the phonemes. While saying the word aloud, children ‘pushed’ each sound into the corresponding box using a penny. The children were then instructed to write the letters corresponding to the phonemes, in each box. Again, while saying the word aloud, children ‘pushed’ each sound into the corresponding box using a penny. The word box procedure was used during the final five minutes of the intervention session.

**Research Design**

This study utilized regression discontinuity to address the research questions, as the design is suitable for the evaluation of programs where participants are selected based on need or merit (Trochim, 2000). The evaluation of school programs is complicated by
several factors, all of which are overcome by use of the regression discontinuity design. First, selection of participants for special programming is not based on random assignment, a requirement for true experimental designs. In many instances certain children meeting certain criteria are deemed eligible to participate in certain programs, and the selection criteria are often related to factors that affect outcomes. Selection criteria are at the heart of the regression discontinuity research design: it assumes that children are not equal on the pre-selection criteria, and is in fact a requirement for the design. Second, the exclusion of children who otherwise qualify for special programming is prohibited by legal, ethical, and political constraints. Randomly excluding children who qualify for the sake of meeting the assumption of random assignment for the true experimental design is often not possible at best, and unethical at worst. Finally, the regression discontinuity research design lends efficiency in the allocation of scare resources by limiting the individuals assigned to the treatment condition to those who need it most (Campbell, 1969).

Overview of Regression Discontinuity

The regression discontinuity research design was first identified as an alternative method of analysis for data obtained via quasi-experimental methods in a research article entitled *Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment*, by Thistlethwaite and Campbell (1960). In this seminal work, the researchers describe a research situation where:
Exposure to an experimental treatment…is determined by the subject’s standing on a single, measured variable, and where the expected effects of the treatment are of much the same nature as would be produced by increasing magnitudes of that variable, examination of the details of the regression may be used to assess experimental effects. The experimental treatment should provide an additional elevation to the regression of dependent variables on the exposure determiner, providing a step-like discontinuity at the cutting score (p. 310).

In their study, Thistlethwaite and Campbell (1960) utilized this method to estimate the motivational effects of social recognition for academic achievement on such variables as: propensity toward earning advanced degrees (MD or PhD), becoming college professors or scientific researchers; and succeeding in obtaining scholarships from scholarship-granting institutions. In the study, two groups of near-winners on the National Merit Scholarship Program received either Certificates of Merit or letters of recommendation, based chiefly on their score on the CEEB Scholarship Qualifying Test (SQT). Those students receiving Certificates of Merit received greater public recognition in the form of newspaper coverage and publication in a booklet distributed to colleges, universities, and other scholarship-granting institutions. Those students receiving letters of recommendation did not receive public recognition. Regressions of each selected variable (e.g., propensity toward earning an advanced degree) on condition of public recognition (ample recognition or no recognition) lines were created. A comparison of the regression lines for the two conditions revealed an additional ‘elevation’, or ‘break’, at the SQT score, suggesting an effect of public recognition. This method of data analysis that refers to the break or discontinuity of the regression line at the cut-off score, was coined the ‘regression discontinuity’ method of analysis.
The regression discontinuity research design is a simple pretest-posttest quasi-experimental research design in which one experimental group is compared to a control group. Groups are selected on the basis of a pre-selected cut-off score on the pre-test measure. The estimation of treatment effect is accomplished by examining the vertical displacement of the intervention group posttest from the pre-post test regression line of the control group: if there exists a ‘disconnect’ or ‘jump’ in the regression line, a treatment effect exists. Using widely accepted notation, the regression discontinuity research design is illustrated as follows:

\[
\begin{array}{cccc}
O & C & X & O \\
O & C & O
\end{array}
\]

where each row references a different group. The ‘O’ refers to the pre- and post-test measurement of each group; the ‘C’ indicates that the groups were assigned using a conditional factor, or selection criteria; and the ‘X’ indicates that a treatment or program was administered.

Logic of Regression Discontinuity

The logic of the regression discontinuity technique is relatively straightforward. If there does not exist a treatment effect for the program or intervention, the relation between the pre-program score and the post-program score would be the same for all students (those who participated in the program and those who did not participate in the
program). The null hypothesis infers that the regression line for those not receiving the intervention is the same for those receiving the intervention (see Figure 3.1). Figure 3.1 presents a hypothetical bivariate distribution of pre- and post-program scores for students. The vertical line demarks the cutoff point; that is, students scoring below the cutoff (to the left of the vertical line) participate in the intervention program; those students scoring above the cutoff (to the right of the vertical line) are in the control, or non-intervention group. Data points to the left of the vertical line represent low scores.

Figure 3.1 Hypothetical Data Demonstrating No Treatment Effect¹ (Trochim, 2006)

on both the pre- and post-measure, and represent those students performing overall, relatively poorly; data points to the right of the vertical line represent relatively higher pre- and post-program scores and represent students performing overall, relatively better. For this example, the cutoff has been arbitrarily set at 50. Those students scoring above 50 on the pre-program measure are considered relatively better functioning and are not in need of treatment or intervention. Those scoring below 50 on the pre-program measure are relatively more poor functioning and are determined in need of treatment. Thus, the data points falling to the right of the cut-off point represent those students who did not receive the treatment or program, and the data points falling to the left of the cut-off represent those students who received the treatment or program. The regression line demonstrates that the relation between the pre- and post-program scores is relatively strong: generally, those who performed low on the pre-program measure performed low on the post-program measure, and those who did well on the pre-program measure performed similarly on the post-program measure. Importantly, this singular regression line predicts the post-program score for all students, without regard to group membership. This reflects a state of ‘non-effect’, where there is no effect due to the intervention or program.

In contrast, Figure 3.2 represents hypothetical data demonstrating an effect due to the intervention or program. For this example, the cut-off score is similarly represented by a vertical line, with the scores for those receiving the intervention falling to the left of the cut-off and the scores for those not receiving the intervention falling to the right of the cut-off. In this hypothetical dataset, the post-program scores for those in the treatment
group have been increased, a reflection of the program effect. The dashed regression line for the treatment group reflects the null hypothesis, and represents the prediction for the post-program scores if the treatment is not effective. The solid regression line for the treatment group represents the prediction for post-program scores based on the observed data. The disconnect, or ‘jump’ in the regression line at the cut-off score reflects that the

![Figure 3.2 Hypothetical Data Demonstrating a Treatment Effect](Trochim, 2006).

Figure 3.2 Hypothetical Data Demonstrating a Treatment Effect² (Trochim, 2006).

post-program scores for those in the treatment group have increased above what was predicted by the original regression line (suggesting no effect due to treatment). Statistical analysis determines whether the difference between the two regression lines is simply due to chance and by how much the two lines differ.

*Regression Discontinuity and True Experiments*

Generally, randomized controlled studies are considered the ‘gold standard’ in research (Linden, Adams, & Roberts, 2004); however, the regression discontinuity method of data analysis for analyzing treatment outcomes has some benefits over randomized controlled trials in the evaluation of program effectiveness. In randomized controlled experiments, requirements for the design include not only random assignment of participants to treatment groups, but also group equivalence on pre-program characteristics. These conditions are required to determine that post-program effects are attributed to the treatment itself. Further, to satisfy the equivalent-groups criteria in randomized controlled designs, some individuals needing the intervention may have services withheld, while other participants may be assigned to receive an unneeded treatment (Lesik, 2006; Doss & Atkins, 2006). The regression discontinuity design is a viable alternative used to make causal inference about the effectiveness of a program when random assignment is unfeasible or unethical. The method by which participants are assigned to the intervention condition guarantees that the groups are different, and builds this difference into the design (Moss & Yeaton, 2006). They are less costly to
conduct, less intrusive, and posses many of the strength of randomized controlled trials (Moss & Yeaton, 2006; Linden, Trochim, & Adams, 2006).

**Assumptions of Regression Discontinuity**

Regression discontinuity requires the following assumptions: (a) the selection of individuals into a treatment group is based on a cut-off score, and is strictly adhered; (b) the relation (e.g., linear, quadratic) between the outcome and selection criteria is known and can be modeled; (c) the changes in the relation between selection and outcome criteria across the groups is due to the treatment; and (d) both groups come from a single pretest distribution (Braden & Bryant, 1990; Trochim, 1984). Each of these assumptions will be discussed in turn.

For regression discontinuity to provide unbiased estimates of a treatment effect, the method of assignment must be based solely on the cut-off score, the integrity of which must be maintained throughout (Lesik, 2006). This assumption is cited in literature as the ‘intention to treat’ principle (Moss & Yeaton, 2006). The cut-off value may be arbitrary and decided by the researcher, or based on professional judgment about which participants might benefit from the program (Shadish, Cook, & Campbell, 2002). Trochim (1984) states that, “The cutoff criterion must be followed. There can be no misassignment relative to the cutoff score” (p 124). Adherence to this assumption is further described by Trochim (1984) as a “major implementation problem for this design” and that “misassignment will result in artificially induced within-group curvilinearity that can lead to biased estimates of treatment effect” (p 153-154). However, some researchers
agree that if the “degree of misassignment is not too great (i.e., if fewer than 5% of the cases are misassigned) the sharp regression-discontinuity analysis will not be seriously biased” (Trochim, 1984, p. 155).

The second assumption of regression discontinuity rests on data fit to a known model, as “one needs to select that subset of variables (i.e., polynomial and interaction terms) that describes the true functional form of the data” (Trochim, 2000, p 128). If the researcher posits a linear model, for example, nonlinear terms including quadratic, cubic, and interaction terms must be ruled out. As stated by Trochim (1984):

The true pre-post test distribution must be describable as a polynomial in x. If the true model is instead logarithmic, exponential, or some other function, the model is misspecified and the estimates of program effect are likely to be inaccurate. If the data can be transformed to a polynomial distribution prior to analysis, the model above may be appropriate. It should be noted that this assumption is not necessarily restrictive. Any functional relationship in the data can be described sufficiently by a high enough order polynomial (p 124).

This model-fitting procedure is known as ‘model specification’, whereby a researcher selects the subset of polynomial and interaction terms that describes the true form of the data. A model might be exactly specified, underspecified, or overspecified. A description of an overspecified and underspecified model is described in Trochim (1984):

A model is overspecified if it includes all variables in the true model along with additional variables that are not in the true model. It is underspecified if it does not include all the terms that are in the true model…parameter estimates will generally be biased when the true model has been underspecified…an overspecified model, on the other hand, will yield an unbiased estimate of program effect, but this estimate will be less efficient (i.e., have greater variance) than one obtained from an exactly specified model (p 129).
In contrast, in an exactly specified model, “the linear and assignment variable terms of the general model are the only ones selected for the analysis, the resulting model will be exactly specified and the estimate of program effect will be unbiased” (Trochim, 1984, p. 129).

The third assumption concerns the implementation of the intervention program. As Trochim (1984) states, “The program conditions must be implemented uniformly within each group. The model assumes that all persons in the program group receive the same ‘amount’ of the program and those in the comparison group receive no program” (p. 124). In this study, the RTI process and Tier 2 intervention services were standardized for first-grade and built into the school’s daily schedule.

The final assumption indicates that both groups must come from a single pretest distribution; that is, the “cutoff value divides this original distribution into two groups” (Trochim 1984, p. 124). This study utilized winter benchmark scores on the DIBELS ORF measure to divide the two groups.

*Internal Validity*

One of the requirements for the use of the regression discontinuity design is that selection is based on a cut-off score. Because there are deliberate pre-program differences between groups, it appears at first glance that selection threats may be problematic. Two selection threats, selection-maturation and selection-regression, have been addressed in the literature (Trochim, 2006). The selection-maturation threat implies
that the difference between the control and program group may simply be due to differing levels of maturation in the two groups. In the regression discontinuity design, a program effect is not indicated by the difference between the post-test means of the two groups, but rather by a change in the pre-post regression line at the cut-off value. Trochim (2006) states, “in order for selection-maturation to be a threat to internal validity in RD designs, it must induce a discontinuity in the pre-post relationship which happens to coincide with the cut-off point – an unlikely scenario in most studies” (Social Research Methods, The Internal Validity of the RD Design, ¶5). The second selection threat, selection-regression, arises when groups are asymmetrically sampled from a distribution. In regression discontinuity designs, the two groups are deliberately asymmetrical, and regression-to-the-mean effects are expected for both groups. That is, we expect those individuals scoring lowest on the pre-program measure will score higher on the post-program measure and those scoring higher on the pre-program measure will score lower on the post-program measure. Trochim (2006) states:

As with selection-maturation, even though we expect to see differential regression to the mean this poses no problem for the internal validity of the RD design. We don’t expect that regression to the mean will result in a discontinuity in the bivariate relationship coincidental with the cut-off point. In fact, the regression to the mean that will occur is expected to be continuous across the range of the pretest scores and is described by the regression line itself (Social Research Methods, The Internal Validity of the RD Design, ¶6).
Power and Effect Size

Power is generally described as the ability of a statistical test to detect a treatment effect if it exists (Gravetter & Wallnau, 2007). For this study, a free computer program, G*Power version 3.0.10, was used to estimate power. Considering 80% power with alpha set at .05, in order to detect a medium treatment effect (0.15) for two predictors the total required sample size is 68. Given the sample size for this study ($n = 141$) the power to detect a small (.02) treatment effect was .30, quite small. However, in regression discontinuity sample size is relatively more important for the control group. In regard to sample size, Trochim (1984) states: “This is more important for the comparison than the program group because it is the comparison group function that describes the null expectation” (p. 124).

While the sample size for the program group may be considered relatively small, it is comparable to other studies that investigate the effects of Tier 2 intervention. For example, a study conducted by Bryant et al, (2008) using regression discontinuity analysis, found medium effects of Tier 2 intervention services in mathematics for 25 students receiving Tier 2 intervention compared to 115 students not receiving Tier 2 intervention. A study investigating the effects of several Tier 2 intervention programs targeting handwriting and composing (Berninger, Rutberg, Abbott, Garcia, Anderson-Youngstrom, Brooks, & Fulton, 2006) found significant effects and contained sample sizes of 14 and 20 for students receiving Tier 2 services. The power analysis for this study revealed in order to detect a treatment effect at 80% power, the magnitude of the
effect size would need to be .07 (somewhere between a small and medium effect) or greater.

Generally, effect sizes reflect the magnitude of mean difference between two groups or strength of an association. Of primary interest for this study is the main effect of program, reflected in the regression slope coefficient for this variable. In regard to regression discontinuity, Trochim (1984) writes, “β₂, the program main effect treatment estimate, is the difference between the two group regression functions at the intercept and at the cutoff” (p 125). The unstandardized regression slope coefficient is an index of effect size based on raw scores and is used when the variables are measured in a meaningful metric. The standardized regression slope coefficient, or ‘beta weight’, is useful when the variables are not measured on a meaningful metric and reflects the influence of a particular variable in terms of standard deviation units. In regard to effect size reporting, Wilkinson (1999) states:

Always present effect sizes for primary outcomes. If the units of measurement are meaningful on a practical level (e.g., number of cigarettes smoked per day) then we usually prefer an unstandardized measure (regression or mean difference) to a standardized measure (r or d) (p 599).

In regard to strength of Keith (2006) in his text *Multiple Regression and Beyond*, proposes the following criteria:

I consider β’s below .05 as too small to be considered meaningful influences on school learning, even when they are statistically significant. β’s above .05 are considered small but meaningful; those above .10 are considered moderate, and those above .25 are considered large (p 62).
The magnitude of effects should be considered within context of a particular area of research (Elmore & Rotou, 2001). Much research to date on specific reading interventions within Tier 2 has used single-subject methodology and measures of effect size appropriate for this type of research, such as percent of non-overlapping data, making comparison of effects difficult. However, a metanalysis of 11 research studies investigating the effects of the RTI approach on student achievement outcomes (Burns, Appleton, and Stehouwer, 2005) and reporting Cohen’s $d$ reveals a mean effect size of .96 for these studies. Differential effects for student achievement were noted for RTI models implemented in the field vs RTI models implemented by faculty for research purposes, with reported mean effect sizes of .62 and 1.23, respectively (Burns et al, 2005). In this study, the main effect of treatment corresponds to the parameter estimate for the program variable and is reported as the unstandardized regression coefficient for that variable. For the purpose of example, a single study (Bryant et al, 2008) investigating the effects of Tier 2 instruction in math using regression discontinuity analysis found no significant effect of Tier 2 instruction for first-grade students ($b = .04$) but a significant effect of Tier 2 instruction ($b = .19$) for second-grade students. According to the criteria proposed by Keith (2006), this is considered a ‘moderate’ or ‘medium’ treatment effect.
The Statistical Model

One of the benefits of regression discontinuity is that it is possible to obtain estimates of program effects if the relationships among the variables are nonlinear (i.e., curvilinear or cubic). This is accomplished using model fitting procedures. As recommended by Trochim (2006), the initial model specified for an analysis contains polynomial terms two orders higher than what is suggested via visual inspection of the data. Although the linear model is often posited, the inclusion of quadratic or cubic terms and their interactions permit the analysis of linearity by determining the non-significance of higher-level terms in the initial model. Under-fitting, or using a linear model when the true form is quadratic or cubic, suggests effects that are not present; over-fitting, or including polynomial terms generally do not suggest such pseudo-effects. With this in mind, the general model for regression discontinuity becomes:

\[ y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + \beta_4 X_i^2 + \beta_5 X_i^2 Z_i + \beta_6 X_i^3 + \beta_7 X_i^3 Z_i + e_i \]

where

- \( y_i \) = outcome for the \( i \)th score,
- \( \beta_0 \) = coefficient for the \( y \)-intercept,
- \( \beta_1 \) = linear coefficient of the transformed pretest,
- \( \beta_2 \) = coefficient for the mean difference between groups,
- \( \beta_3 \) = linear interaction coefficient,
- \( \beta_4 \) = quadratic transformed pretest coefficient,
- \( \beta_5 \) = quadratic interaction coefficient,
- \( \beta_6 \) = cubic transformed pretest coefficient,
- \( \beta_7 \) = cubic interaction coefficient,
- \( X_i \) = transformed pretest: \( x_i - (\text{cutpoint}) \),
- \( Z_i \) = Dummy variable for intervention (0 = comparison and 1 = treatment), and
- \( e_i \) = residual for the \( i \)th score.
It is important to note that \( X_i \), (the pretest) is the *transformed pretest*; that is, the transformed pretest represents each score on the pretest, minus the cutoff value used to create the groups. Transforming the pretest in this manner is important, as:

> This has the effect of moving the cutoff value to the y-intercept point and improves the interpretability of the coefficients...when this transformation is applied, the \( \beta_0 \) term is both the y-intercept for the comparison group regression line and the predicted y-value for the comparison group at cutoff. From this, it follows that \( \beta_2 \), the program effect estimate, is the difference between the two group regression functions at the intercept and at the cutoff (Trochim, 1984, p. 125).

For this study, a linear model is posited. Following Trochim’s (1984) recommendation to include polynomials two orders higher than what appears via visual inspection of the data, quadratic terms were included in the initial model. Following this, the initial model for this study becomes:

\[
y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + \beta_4 X_i^2 + \beta_5 X_i^2 Z_i + e_i
\]

where

\[
y_i = \text{spring ORF score for the } i\text{th score,}
\]
\[
\beta_0 = \text{coefficient for comparison group y-intercept, at cutoff}
\]
\[
\beta_1 = \text{linear coefficient of the transformed winter ORF score (pretest),}
\]
\[
\beta_2 = \text{coefficient for the program effect estimate,}
\]
\[
\beta_3 = \text{linear interaction coefficient,}
\]
\[
\beta_4 = \text{quadratic transformed pretest coefficient,}
\]
\[
\beta_5 = \text{quadratic interaction coefficient,}
\]
\[
X_i = \text{transformed pretest: } x_i - \text{(cutpoint)},
\]
\[
Z_i = \text{Dummy variable for intervention (0 = comparison and 1 = treatment),}
\]
\[
e_i = \text{residual for the } i\text{th score.}
\]

For this analysis, the null hypothesis of interest is:

\[
H_0: \beta_2 = 0
\]
That is, the mean difference between the intervention and control groups at the cutoff is zero. The alternative hypothesis becomes:

\[ H_0: \beta_2 \neq 0 \]

That is, the mean difference between the intervention and control groups at the cutoff is not zero.

**Description of Program Fidelity**

Information regarding the nature, frequency and duration of the intervention procedure was provided to the researcher by the district’s school psychologist, who collected information from the intervention teachers on the researcher’s behalf. Questions were open-ended, and were designed to determine not only the consistency in

<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Typically, how long were the individual intervention sessions?</td>
</tr>
<tr>
<td>2. How many students did you see at one time?</td>
</tr>
<tr>
<td>3. Where did you administer the interventions?</td>
</tr>
<tr>
<td>4. For how many weeks did the intervention last?</td>
</tr>
<tr>
<td>5. Did you modify the intervention program in any way?</td>
</tr>
<tr>
<td>6. Were there any difficulties you encountered while conducting the</td>
</tr>
<tr>
<td>interventions?</td>
</tr>
<tr>
<td>7. Do you feel that students progressed over the course of the</td>
</tr>
<tr>
<td>intervention? Why or why not?</td>
</tr>
</tbody>
</table>

Figure 3.3 Questions Assessing Program Fidelity
the level of implementation between the intervention teachers, but also to obtain the teachers’ perceptions about the effectiveness of the intervention. Figure 3.3 provides a listing of questions that were asked of the intervention teachers. The district school psychologist distributed the questions and collected the responses. Responses were then provided to the researcher. Once obtained, the teachers’ responses were compared to determine the consistency in which interventions were implemented. Information regarding barriers to intervention implementation and perceptions about the students’ progress were examined to determine whether any themes emerged.

This chapter provided a discussion of the source of data, measures used to track progress, and the intervention program used by the school district. A detailed discussion of the research design, along with the statistical model, was also presented. Chapter 4 will provide detailed steps used for conducting regression discontinuity analysis and the quantitative and qualitative results.
Chapter 4

Results

This chapter presents the quantitative findings obtained through analysis using regression discontinuity and qualitative information, provided by the teachers through the district school psychologist. Before presenting the quantitative findings, the method of analysis is outlined. The steps required to analyze data from a regression discontinuity design are described by Trochim (2006) and are presented here.

Steps for Conducting Regression Discontinuity

Step 1: Transform the Pretest

The pretest variable must be transformed by subtracting the cutoff value from each pretest score. When this is done, the intercept becomes equal to the cutoff value on the pretest. Trochim (2006) states:

If we subtract the cutoff from every pretest value, the modified pretest will be equal to zero, where it was originally at the cutoff value. Since the intercept is by definition the y-value when x = 0, what we have done is set x to 0 at the cutoff, making the cutoff the intercept point (Social Research Methods, Analysis Section, ¶ 1).

In statistical notation, the transformation becomes:

\[ \tilde{X} = X_i - X_c \]

where
\( \tilde{X} = \) the transformed pretest, \\
\( X_i = \) the original pretest value for the ith person, and \\
\( X_c = \) the cutoff value

Step 2: Examine the relationship visually

The second step is to create a bivariate plot, or graph of the pre-post test relationship. This allows the researcher to determine whether there is a visible discontinuity in the graph at the cutoff value. Trochim (2006) states:

The discontinuity could be a change in level vertically (main effect), a change in slope (interaction effect) or both. If it is visually clear that there is a discontinuity at the cutoff, then one should not be satisfied with analytic results which indicate no program effect. However, if no discontinuity is visually apparent, then it may be that variability in the data is masking an effect and one must attend carefully to the analytic results (Social Research Methods, Analysis Section, ¶ 2).

Examination of the bivariate plot also permits the researcher to determine the degree of polynomial that may need to be included in the analysis. As Trochim (2006) recommends:

A good approach is to count the number of flexion points (i.e., number of times the distribution ‘flexes’ or ‘bends’) which are apparent in the distribution. If the distribution appears linear, there are no flexion points. A single flexion point could be indicative of a second (quadratic) order polynomial. This information will be used to determine the initial model which will be specified (Social Research Methods, Analysis Section, ¶ 3).
Step 3: Specify Higher-Order Terms and Interactions

The third step involves transforming the modified pretest variable. After examining the number of flexion points in the bivariate graph, the researcher creates transformations two orders of polynomial higher than the number of flexion points suggested in the previous step. For example, if the data appears linear, the researcher would create transformations up to a second order \((0 + 2)\) polynomial Trochim (2006) explains:

The first order polynomial already exists in the model \((X)\) and so one would only have to create the second order polynomial by squaring \(X\) to obtain \(X^2\). For each transformation of \(X\) one also creates the interaction term by multiplying the polynomial by \(Z\)…each transformation can be easily accomplished through straightforward multiplication on the computer (Social Research Methods, Analysis Section, ¶ 5).

Trochim (2006) further recommends, as a rule of caution, “overestimating the true polynomial function” by including all higher-level terms and their associated interactions (Social Research Methods, Analysis Section, ¶ 5).

Step 4: Estimate the Initial Model

In this step, the researcher is ready to estimate the initial model by using a computer program capable of running multiple regression analyses. In this step, the researcher “simply regresses the posttest score, \(Y\), on the modified pretest \(X\), the treatment variable \(Z\), and all higher-order transformations and interactions created in Step 3” (Trochim, 2006, Social Research Methods, ¶ 6). The regression coefficient of the membership, or program variable, reflects the main effect of program. If there exists a vertical
discontinuity at the cutoff value, it will be reflected in the estimation of this coefficient. The significance of the regression coefficient is tested by constructing a $t$ test using the standard error of the regression coefficient. Trochim (2006) states:

If the analyst at Step 3 correctly overestimated the polynomial function required to model the distribution then the estimate of the program effect will at least be unbiased. However, by including terms which may not be needed in the true model, the estimate is likely to be inefficient, that is, standard error terms will be inflated and hence the significance of the program effect may be underestimated. Nevertheless, if at this point in the analysis the coefficient is highly significant, it would be reasonable to conclude that there is a program effect (Social Research Methods, Analysis Section, ¶ 7).

Step 5: Refine the Model

As Trochim (1984) states, “the primary goal in regression discontinuity is to obtain an unbiased and statistically efficient treatment effect. This is perfectly achieved only if the subset of variables that is selected from the general model exactly describes the true model” (p 128). Model fitting is a “tricky procedure and should be approached cautiously if one wishes to minimize the possibility of bias” (Trochim, 2006, Social Research Methods, Analysis Section, ¶ 8). Generally, model fitting requires the researcher to examine the degree to which the overall model fits the data, the presence of insignificant coefficients, and the pattern of residuals (errors). Trochim (2006) advises:

A conservative way to decide how to refine the model would begin by examining the highest-order term in the current model and its interaction. If both coefficients are non-significant, and the goodness-of-fit measures and pattern of residuals indicate a good fit one might drop these two terms and re-estimate the resulting model. Thus if one estimated up to a fourth-order polynomial, and found the coefficients for $X^4$ and $X^4Z$ were non-significant, these terms can be dropped and
the third-order model respecified. One would repeat this procedure until: a) either of the coefficients is significant; b) the goodness-of-fit drops appreciably; or, c) the pattern of residuals indicates a poorly fitting model. The final model may still include some unnecessary terms, but there are likely to be fewer of these and consequently, efficiency should be greater (Social Research Methods, Analysis Section, ¶ 8).

Researchers might be tempted to include higher-order terms to ensure that the estimates are unbiased. In addressing this, Trochim (1984) advises:

As more terms are added to the model the potential for bias is reduced, but so is the precision of the estimates. Furthermore, higher-order models will ultimately fit the data well, but will often yield functions that are clearly absurd in any substantive sense. Typically, models higher than second or third-order will not be theoretically justifiable in most social science arenas...deliberate over-fitting of the ‘likely’ true function will tend to lead to unbiased estimates that are somewhat imprecise (p 133).

Step 6: Determine Effects

While research reports often cite level of significance, or p values for hypothesis tests, tests of statistical significance do not “provide any real information about the absolute size of a treatment effect” (Gravetter & Wiallnau, 2007, p 256). Volker (2006) notes:

The p value is often discussed as if it is an index of effect size or meaningfulness of the result. It is not unusual to read in an article something like, ‘the result was highly significant (p < .0001),’ which a large sample size can lead to a small p value for even a minuscule effect size makes the p value useless for this purpose (p 654).

In this study, the main effect of treatment corresponds to the parameter estimate for the program variable and is reported as the standardized regression coefficient for that
variable. Much research to date on specific reading interventions within Tier 2 has used single-subject methodology and measures of effect size appropriate for this type of research, such as percent of non-overlapping data, making comparison of effects difficult. However, a study conducted by Hagans (2008) which investigated the effects of a small group reading intervention for first-grade students and used the DIBELS Nonsense Word Fluency as the outcome variable, found moderate to large effects ($\beta = .084$) indicating that students in the intervention group experienced an average increase of 15 words per minute. Another study (Marchand-Martella, Martella, Kolts, Mitchell, & Mitchell, 2006) using DIBELS fluency subtests for pre- and post-intervention measures found statistically significant findings for first grade reading intervention, with Cohen’s $d$ effect sizes ranging from 2.42 (or about a 14 words per minute increase) for Letter Naming Fluency to 1.36 (or about 31 words per minute) for Nonsense Word Fluency.

Results for Regression Discontinuity

Table 4.1 presents descriptive statistics for the winter and spring ORF scores, and includes statistics describing skew and kurtosis. Examination of the coefficients for skew and kurtosis reveal non-normal distributions. In regression “multivariate normality is not required when estimating function coefficients of structure coefficients (i.e., parameter estimation), but when evaluating the results of the multivariate analyses, the underlying assumption is that the distributions are normal” (George, 2001, p. 4). Frequently, outliers, or extreme data points are the culprit but their removal from the analysis should be done with caution, as they may represent a ‘true’ state of affairs in the data (Keith,
Other potential causes of non-normality include errors in data entry or nondeclared missing values (Osbourne, 2008). Prior to running the analysis, the dataset was reviewed for data-entry errors and potential outliers. While outliers were present, they were not removed prior to the analysis as they are considered valid data points.

Table 4.1 Descriptive Statistics for First-grade Winter and Spring ORF

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>S.D.</th>
<th>Skew (S.E)</th>
<th>Kurtosis (S.E.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervention</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winter</td>
<td>21</td>
<td>4.81</td>
<td>5.00</td>
<td>2.38</td>
<td>-0.24 (0.50)</td>
<td>-1.14 (0.97)</td>
</tr>
<tr>
<td>Spring</td>
<td>21</td>
<td>22.48</td>
<td>22.00</td>
<td>12.08</td>
<td>0.46 (0.50)</td>
<td>0.97 (-0.59)</td>
</tr>
<tr>
<td>Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winter</td>
<td>121</td>
<td>39.86</td>
<td>28.00</td>
<td>30.03</td>
<td>1.12 (0.22)</td>
<td>0.23 (0.44)</td>
</tr>
<tr>
<td>Spring</td>
<td>121</td>
<td>66.36</td>
<td>60.00</td>
<td>34.67</td>
<td>0.49 (0.22)</td>
<td>-0.68 (0.44)</td>
</tr>
</tbody>
</table>

Important for regression is the assumption of normality of the error distribution; that is, the errors or residuals themselves are normally distributed (Keith, 2006). One possible cause of normality violations is that the distributions of the dependent and/or independent variables are significantly non-normal (Nau, 2005, Testing the Assumptions of Linear Regression, ¶ 17). A p-p plot provides one method for diagnosing violations of normality of the error distribution, and is a plot of the cumulative area under the normal curve for the observations plotted against the cumulative area under the normal curve for the expected values. The residuals are considered normally distributed if they fall on or close to the diagonal line. If a bow-shaped pattern or S-shaped pattern is observed, this
indicates that the assumption of normality of the residuals has been violated. Because the
distribution of variables reveals some skew and kurtosis which can in turn affect this
assumption and thus the estimation of the coefficients, a p-p plot was constructed for
each model run using the raw (untransformed data). The results of the p-p plots revealed
some departure from normality for the raw ORF scores. One possible cause of normality
violations is that the distributions of the dependent and/or independent variables are
significantly non-normal (Nau, 2005, Testing the Assumptions of Linear Regression, ¶ 17) and a possible remedy involves transformation of data. “For methods that rely on
normality of the data, direct manipulation of the data to make the random errors
approximately normal is usually the best way to bring the data in line with this
assumption” (NIST/SEMATECH, 2006, Accounting for Errors with a Non-Normal
Distribution, ¶ 1). Given that there is some departure from normality of the residuals for
all models using non-transformed ORF scores, the researcher decided to conduct the
regression discontinuity analysis after applying transformations to the data. The process
for transforming data is described as follows:

1. Transform the response variable to make the distribution of the random errors
   approximately normal.
2. Transform the predictor variables, if necessary, to attain or restore a simple
   functional form for the regression function.
3. Fit and validate the model in the transformed variables.
4. Transform the predicted values back into the original units using the inverse of the transformation applied to the response variable (NIST/SEMATECH, 2006, Accounting for Errors with a Non-Normal Distribution, ¶ 2)

Following the recommendation of Cleveland (1984) transformation using base 10 logs was selected for this study as “in cases where there are extremes of range, base 10 is desirable” (Osbourne, 2008, p 3). In brief, a logarithm is the power (exponent) to which a base must be raised in order to get the original number (Osbourne, 2008). For example, if a log with a base of 10 is used, then \( \log_{10}(100) = 2 \); that is, the number 10 must be raised to the second power to yield the number 100.

Table 4.2 provides descriptive statistics for the winter and spring ORF scores, transformed using base 10 logarithms. Table 4.3 provides a comparison of the index of skew and kurtosis for the raw (untransformed) and log transformed variables. The skew and kurtosis indices were computed by dividing the value of skew/kurtosis by its respective standard error. Standard normal distributions have

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>S.D.</th>
<th>Skew (S.E)</th>
<th>Kurtosis (S.E)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Intervention</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winter</td>
<td>5.46</td>
<td>5.50</td>
<td>2.83</td>
<td></td>
<td>-0.01 (0.47)</td>
<td>-0.93 (0.92)</td>
</tr>
<tr>
<td>Spring</td>
<td>22.33</td>
<td>20.50</td>
<td>11.19</td>
<td></td>
<td>0.49 (-0.55)</td>
<td>0.47 (0.92)</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Winter</td>
<td>40.62</td>
<td>29.00</td>
<td>30.03</td>
<td></td>
<td>1.10 (0.22)</td>
<td>1.71 (0.44)</td>
</tr>
<tr>
<td>Spring</td>
<td>67.50</td>
<td>64.00</td>
<td>34.32</td>
<td></td>
<td>0.48 (0.22)</td>
<td>-0.68 (0.44)</td>
</tr>
</tbody>
</table>
skewness and kurtosis of zero. Indcies that are greater or less than zero represent
distributions that depart from normality. The results in Table 4.3 reveal that the data
transformation resulted in slight improvements in skew and kurtosis for the winter and
spring administration of the ORF scores for both the intervention and control group, with
one exception. The index of kurtosis for the control group, winter administration is .52
for the raw data vs 3.88 for the transformed data.

Table 4.3 Skew and Kurtosis Comparisons

<table>
<thead>
<tr>
<th>Group</th>
<th>Skew (Raw)</th>
<th>Skew (Transformed)</th>
<th>Kurtosis (Raw)</th>
<th>Kurtosis (Transformed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervention</td>
<td>Winter</td>
<td>-.48</td>
<td>-.02</td>
<td>-1.17</td>
</tr>
<tr>
<td></td>
<td>Spring</td>
<td>.92</td>
<td>-.89</td>
<td>-1.64</td>
</tr>
<tr>
<td>Control</td>
<td>Winter</td>
<td>5.09</td>
<td>5.00</td>
<td>.52</td>
</tr>
<tr>
<td></td>
<td>Spring</td>
<td>2.23</td>
<td>2.18</td>
<td>-1.54</td>
</tr>
</tbody>
</table>

Prior to running the regression discontinuity analysis, data codes were constructed
to ease interpretation. Table 4.4 presents the numerical codes pertaining to data in the
analysis tables.
Table 4.4 Codes for Data Analysis

<table>
<thead>
<tr>
<th>Column Heading</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogWi</td>
<td>log 10 of the assignment variable score, Winter ORF</td>
</tr>
<tr>
<td>LogSpr</td>
<td>log 10 of the post-program score, Spring ORF</td>
</tr>
<tr>
<td>PreCut</td>
<td>transformed assignment variables score</td>
</tr>
<tr>
<td>Group</td>
<td>group membership</td>
</tr>
<tr>
<td>Interact</td>
<td>linear interaction term</td>
</tr>
<tr>
<td>Quad</td>
<td>quadratic term</td>
</tr>
<tr>
<td>QuadInt</td>
<td>quadratic interaction term</td>
</tr>
</tbody>
</table>

The first step in the analysis was to transform the assignment, or pre-test variable (here, the winter administration of ORF) using the cutoff score. The cutoff score on the assignment variable was established as 8 words read correctly on the winter administration of the DIBELS ORF. This is accomplished by subtracting the cutoff value from each score to transform the pretest scores. Transforming the assignment variable in this fashion essentially sets the cutoff score equal to the intercept point (Trochim, 2006). Since it was necessary to transform the scores into logs, the transformation of the assignment variable was conducted in two steps. First, the assignment variable was recalculated using logs. Then, the cutoff value in logs was subtracted from each score to transform the pretest scores. Transforming the assignment variable in this fashion essentially sets the cutoff score equal to the intercept point (Trochim, 2006).

Participation in Tier 2 services was denoted by the group membership variable, where: 1 = intervention services, and 0 = no intervention services (control group).

The second step in the analysis required the researcher to visually examine the relationship using a bivariate scatterplot (see Figure 4.1). In this scatterplot, the ‘dashed’ regression line represents the null condition, or the predicted performance for students in
the absence of the treatment (here, Tier 2 services). The ‘solid’ regression line represents the predicted performance for students under the treatment condition. A treatment effect is suggested if there appears to be a ‘disconnect’ or ‘discontinuity’ between the two regression lines at cutoff. The bivariate scatterplot shown in Figure 4.1 appeared to reveal a change or ‘jump’ in the regression lines at the cutoff point, which suggested a program effect. Although there did not appear to be flexion points (‘bends’ in the distribution), it was difficult to ascertain if this indeed was the case. Taking a conservative approach and

Figure 4.1 Bivariate distribution of spring ORF and assignment variable
to ensure that the model was not underspecified and lacking important terms in the model, the researcher included the linear interaction, quadratic, and quadratic interaction terms in the initial model fitting procedure.

Table 4.5 provides the results of the initial model specification, incorporating the linear and quadratic terms and their associated interactions. Results indicate that as a whole, the model explains or predicts 81.7% of the variance in the log of spring ORF scores, $F(136, 5) = 121.260, \ p < .001, \ R^2 = .817$. The primary effect of interest is the treatment effect estimate for the ‘group’ variable, which is $1.78 (SE .071)$ for this model.

<table>
<thead>
<tr>
<th>Model</th>
<th>$B$</th>
<th>Std. Error</th>
<th>Beta</th>
<th>t</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>1.250</td>
<td>.047</td>
<td></td>
<td>26.618</td>
<td>.000</td>
</tr>
<tr>
<td>Precut</td>
<td>1.115</td>
<td>.175</td>
<td>1.610</td>
<td>6.369</td>
<td>.000</td>
</tr>
<tr>
<td>Group</td>
<td>.178</td>
<td>.071</td>
<td>.206</td>
<td>2.488</td>
<td>.014</td>
</tr>
<tr>
<td>Interact</td>
<td>-.789</td>
<td>.391</td>
<td>-.397</td>
<td>-2.018</td>
<td>.046</td>
</tr>
<tr>
<td>Quad</td>
<td>-.311</td>
<td>.137</td>
<td>.432</td>
<td>-2.416</td>
<td>.017</td>
</tr>
<tr>
<td>Quadint</td>
<td>.012</td>
<td>.399</td>
<td>.005</td>
<td>.029</td>
<td>.977</td>
</tr>
</tbody>
</table>

Note. Dependent Variable DIBELS ORF, Log Transformed

$R = .904 \quad R^2 = .817 \quad R^2_{(adj)} = .810$

Examination of this model reveals that the quadratic interaction term (Quadint) was not significant, and could be eliminated from the model. Table 4.6 provides the results of
model 2, the model with the quadratic interaction term (Quadint) eliminated. Overall, this model explains or predicts 81.7% of the variance in the outcome variable, \( F(137, 4) = 152.688, \ p < .001, \ R^2 = .817. \)

Table 4.6  Regression Results for Model 2 (Without Quadratic Interaction Term)

<table>
<thead>
<tr>
<th>Model</th>
<th>( B )</th>
<th>Std. Error</th>
<th>Beta</th>
<th>( t )</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>1.250</td>
<td>.045</td>
<td></td>
<td>27.906</td>
<td>.000</td>
</tr>
<tr>
<td>Precut</td>
<td>1.113</td>
<td>.164</td>
<td>1.608</td>
<td>6.774</td>
<td>.000</td>
</tr>
<tr>
<td>Group</td>
<td>.176</td>
<td>.055</td>
<td>.205</td>
<td>3.226</td>
<td>.002</td>
</tr>
<tr>
<td>Interact</td>
<td>-.789</td>
<td>.294</td>
<td>-.400</td>
<td>-2.706</td>
<td>.008</td>
</tr>
<tr>
<td>Quad</td>
<td>-.329</td>
<td>.128</td>
<td>-.430</td>
<td>-2.572</td>
<td>.011</td>
</tr>
</tbody>
</table>

Note. Dependent Variable DIBELS ORF, Log Transformed

\[ R = .904 \quad R^2 = .817 \quad R^2_{(adj)} = .811 \]

Again, the primary effect of interest is the treatment effect estimate for the ‘group’ variable, which for model 2 is .176 (SE .055). Elimination of the quadratic interaction term in model 2 yielded a slightly improved parameter estimate for the group, or treatment variable. This is reflected in the smaller standard error in model 2. Parameter estimates for the linear interaction term (Interact) and the quadratic term (Quad) were determined significant and indicated to the researcher that these terms should be retained.
in the model. To aid in this decision, a bivariate scatterplot incorporating parameter estimates from model 2 was created and is presented in Figure 4.2.

![Bivariate distribution of spring ORF and assignment variable, quadratic](image)

Figure 4.2 Bivariate distribution of spring ORF and assignment variable, quadratic

In this figure, the ‘dashed’ line represents the null condition, or the predicted scores for individuals not receiving Tier 2 services. The ‘solid’ line represents the predicted scores for those individuals receiving Tier 2 services. Visual inspection of the fit of the data points to the regression lines reveals what appears to be a ‘jump’ or discontinuity in the lines at cutoff. Further, the difference in slope of the tangent line seems to provide some support for an interaction.
For the sake of comparison, a third model, which eliminated all quadratic terms, was analyzed. These results are presented in Table 4.7.

Table 4.7 Regression Results for Model 3 (Without Quadratic Term)

<table>
<thead>
<tr>
<th>Model</th>
<th>B</th>
<th>Std. Error</th>
<th>Beta</th>
<th>t</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>1.345</td>
<td>.026</td>
<td></td>
<td>51.710</td>
<td>.000</td>
</tr>
<tr>
<td>Precut</td>
<td>.702</td>
<td>.039</td>
<td>1.014</td>
<td>17.943</td>
<td>.000</td>
</tr>
<tr>
<td>Group</td>
<td>.112</td>
<td>.050</td>
<td>.130</td>
<td>2.256</td>
<td>.026</td>
</tr>
<tr>
<td>Interact</td>
<td>-.091</td>
<td>.109</td>
<td>-.046</td>
<td>-.836</td>
<td>.404</td>
</tr>
</tbody>
</table>

Note. Dependent Variable DIBELS ORF, Log Transformed

R = .899 \[R^2 = .808\] \[R^2_{(adj)} = .804\]

The $R^2$ value for model 3 is .808, indicating that model 3 explains, or predicts 80.8% of the variance in log spring ORF scores; while this is a slight reduction, the difference in the percent of variance explained by model 2 verses model 3 is relatively small (0.9%). The treatment effect estimate for the ‘group’ variable in model 3 is .112 ($SE .050$). The reduction in standard error for the ‘group’ variable suggests a slightly improved parameter estimate. Further review of the results of model 3 suggests that the interaction term is non-significant. A final model, with the linear interaction term eliminated, was run to determine how well a linear model describes the data. The results of the final model, model 4, is presented in Table 4.8.
Table 4.8  Regression Results for Model 4 (Without Linear Interaction Term)

<table>
<thead>
<tr>
<th>Model</th>
<th>B</th>
<th>Std. Error</th>
<th>Beta</th>
<th>t</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>1.352</td>
<td>.025</td>
<td></td>
<td>54.856</td>
<td>.000</td>
</tr>
<tr>
<td>Precut</td>
<td>.690</td>
<td>.036</td>
<td>.997</td>
<td>18.930</td>
<td>.000</td>
</tr>
<tr>
<td>Group</td>
<td>.128</td>
<td>.045</td>
<td>.149</td>
<td>2.823</td>
<td>.005</td>
</tr>
</tbody>
</table>

Note. Dependent Variable DIBELS ORF, Log Transformed

R = .898  \(R^2 = .807\)  \(R^2_{(adj)} = .804\)

Analysis of the results of model 4 reveals that this model explains or predicts 80.7% of the variance in log spring ORF scores. While this represents a slight reduction when compared to model 3, the difference (0.1%) is extremely small. The treatment effect estimate for the ‘group’ variable in model 4 is .128 (SE .045), which is slightly greater than that for model 3. The reduction in standard error for the ‘group’ variable suggests a slightly improved parameter estimate.

Trochim (1984) makes the following recommendation in model selection and interpretation of treatment effects:

If the evidence is compelling, the analyst can be satisfied to accept the likely function as an appropriate model. In the absence of such evidence, it is generally preferable to report multiple estimates, or if a single estimate is required, to select one from a reasonable higher-order step in the analysis (p 132).

In model selection, the goal is to select a model that is exactly specified, and if not, then overspecified. Taking a conservative approach, the treatment effect estimates from
Model 2, which is perhaps exactly specified, and Model 4, which is perhaps overspecified, are selected and reported. The equation for Model 2 (the model with the linear interaction and quadratic term) becomes:

\[ y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + \beta_4 X_i^2 + e_i \]

The null hypothesis of interest was \( \beta_2 = 0 \), with the alternative hypothesis of \( \beta_2 \neq 0 \). Given this formula, \( \beta_2 = .176 \). The equation for Model 4 (the model excluding the linear interaction and all quadratic terms) becomes:

\[ y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + e_i \]

The null hypothesis of interest was \( \beta_2 = 0 \), with the alternative hypothesis of \( \beta_2 \neq 0 \). Given this formula, \( \beta_2 = .128 \) Using interpretive guidelines presented by Keith (2006) standardized beta coefficients consistently reveal a ‘medium’ treatment effect for Model 4 and Model 2. The implications of the treatment effect are reported in Table 4.9, which provides a description of students according to classification of risk. On the winter administration of the ORF measure, 18 students fell in the ‘at-risk’ category; 45 students fell in the ‘some risk’ category; and 78 fell in the ‘low risk’ category. On the spring administration of the ORF measure, 14 students fell in the ‘at-risk’ category; 53 students fell in the ‘some risk’ category; and 74 fell in the ‘low risk’ category.
Table 4.9 Benchmark Scores Corresponding to Classification of Risk, by Administration

<table>
<thead>
<tr>
<th>ORF Risk Category</th>
<th>Winter N=141</th>
<th>Spring N=141</th>
</tr>
</thead>
<tbody>
<tr>
<td>At Risk</td>
<td>18</td>
<td>14</td>
</tr>
<tr>
<td>Some Risk</td>
<td>45</td>
<td>53</td>
</tr>
<tr>
<td>Low Risk</td>
<td>78</td>
<td>74</td>
</tr>
</tbody>
</table>

These findings reveal a 22% reduction in the number of students classified as ‘at risk’ on the spring assessment; however, it should be noted that some of those students included in the ‘at risk’ category on the spring assessment were students who participated in the control group and were originally in a ‘lower’ risk category on the winter assessment.

**Qualitative Information**

Qualitative information was provided to the researcher through the district school psychologist. Reports regarding the nature of the Tier 2 intervention were consistent across both intervention specialists. In regard to the time allotted for the intervention, both teachers were fairly consistent in their reports. One teacher reported, “we schedule about 30 minutes, but by the time you really sit down with them I would put 25 minutes, 5 times per week”. This is consistent with the second intervention teacher, who reports, “All intervention groups are 30 minutes long (due to scheduling constraints)”. Both teachers report that intervention groups are small in number. One teacher stated, “we limit our groups to no more than six kids per group-especially with the ‘yellow’ and ‘red’ kids (i.e., Tier 3 and Tier 2 students), for more individualized instruction”. The second intervention teacher stated, “we generally have six kids per group, but this can change
depending upon student absences, of course”. Both teachers reported that the intervention sessions were conducted in a small room, outside of the regular classroom setting. In regard to the content of the Tier 2 intervention, both teachers were consistent in the reports. One intervention teacher indicated that, “everyone gets small group reading instruction using the program five times per week no matter what reading level they are on” and that the intervention consists of “reading of instructional leveled text (through Early Success series), making words (phonemic awareness and phonics activity) by Elkonin boxes (push a penny into each box when you hear a different sound)”. The other intervention teacher reported that by using the program, “we focus on reading comprehension, phonics skills, and how to answer short extended-response questions – like the ones they’ll see on the OAT type tests”. Both teachers reported that the intervention continued from the winter benchmark administration to the spring benchmark administration, or “about 10 weeks”.

The second goal of the qualitative measure was to determine whether any themes emerged regarding perceived barriers or factors that may have affected the students’ level of success during the Tier 2 intervention. Teachers consistently reported two factors that they felt influenced the effectiveness of Tier 2 intervention. The factors seemed to fall within two categories: time constraints and student characteristics. One teacher candidly reported on this by stating, “We only have 30 minutes with these kiddos and try to shove as much learning into them as possible in that time frame”. Both teachers were consistent in their perceptions that for many students, student behavior and level of cognitive ability were factors that impeded performance. One teacher reported that, “Students move up but
are still not on grade level. I do not know if they will ever really be able to catch up with even strong intervention because focusing is an issue and level of cognitive ability”. The second intervention teacher cites one particular student as an example: “right now I work with one ‘on his way’ (to being evaluated for special education) student for 25 minutes, five times per week, and do leveled reading with him after lunch to see if he’s moved up in reading at all. He has, but is still not on grade level. He seems really low in ability”.

The district school psychologist reflected on this, by stating, “one of the reasons we use the 10th instead of the 15th percentile to select students for Tier 2 is that there are so many kids who fall in this category – so many kids who are just basically ‘lower-functioning’.”

83
Chapter 5
Discussion

The purpose of this study was twofold: to determine the effects of Tier 2 instruction in reading and to demonstrate the utility of the regression discontinuity design when analyzing RTI data. This chapter will reflect on the findings as they correspond to the research questions.

Summary of Findings

Effects of Tier 2 Instruction in Reading

Results of this study yielded standardized effects sizes of .205 for Model 2 (which, in addition to the pre-test and group variable, also included the linear interaction and quadratic term) and .149 for Model 4 (which includes only the pre-test and group variables). According to the criteria proposed by Keith (2006) and reported in Chapter 4, these treatment effects are considered ‘medium’ effects. Examination of student scores according to classification of risk (see Table 4.9) reveals a 22% reduction in the number of students classified as ‘at risk’ on the spring administration of the ORF measure of the DIBELS. Since the overarching goal is to increase students’ achievement, this is an encouraging finding. While it appears that program effects were found to be significant, the skewed nature of the data required transformation, which makes interpretation
difficult. As stated by Osbourne (2008), “the very act of altering the relative distances between data points, which is how these transformations improve normality, raises issues in the interpretation of the data” (p. 4). Given this, the obtained results should be interpreted with caution.

In *A New View of Statistics*, Hopkins (2008) indicates that log-transformed data is often interpreted in terms of percentages; that is, the findings represent a percent change in the dependent variable with the associated increase in the independent variable. Information provided on the UCLA Academic Technology Services website ([http://www.ats.ucla.edu](http://www.ats.ucla.edu)) indicates that log-transformed data is interpreted in the following way: “…the format for interpretation is that dependent (sic) variable changes by 100*(coefficient) percent for a one unit increase in the independent variable while all other variables in the model are held constant” (Section, ‘How can I Interpret Log Transformed Variables in Linear Regression?’, ¶ 3).

For this study, the independent variable under investigation was the effect of program, or Tier 2 services. Applying the interpretive guidelines provided on the UCLA website, and applying the effects found in this study, this means that a student receiving Tier 2 intervention services will experience a ‘maximum’ effect of a 17.6 percent increase in the log of the spring score, or a ‘minimum’ effect of 12.8 percent increase in the log of the spring score. A student performing at cut-off (or eight words per minute, for this study) on the pre-program administration would predictably score 11.22 or 10.23 points higher, respectively, on the next administration of the ORF of the DIBELS. This finding is consistent with previous studies (Hagans, 2008; Marchand-Martella et al, 2006)
which report similar gains using DIBELS fluency measures. For this particular sample of students, this means that a student performing at cut off would move into a lower category of risk according to DIBELS classification.

One of the benefits of using regression discontinuity is that the interaction between the pre-test score and treatment can be determined, as well as non-linear trends. While these are not the primary effects of interest in this study, it is important to reflect on the findings. One model (Model 2) examined in this study yielded a significant linear interaction effect and a significant quadratic trend. Interaction effects indicate “whether the intervention is particularly effective with a subgroup of the students who received intervention (typically, those who score highest or lowest on the pretest)” (Bryant, et.al., 2008). In this study, Model 2 yielded a significant linear interaction, which suggests that the effects of the program differed depending on the level of the pretest. The quadratic term was also significant and negative in valence, suggesting a trend that is ‘downward’, or concave in shape. The graph in Figure 4.2 shows that the lower performing students in the Tier 2 intervention group demonstrated more improvement in their rate of acquisition than did higher performing students in the intervention group. This is an interesting finding, which may reflect either: (a) a ‘floor effect’ for those scoring lowest (they had nowhere to go but up), (b) the intervention did not teach higher performing students skills beyond which they already possessed, or (c) a qualitative difference in subgroups of children determined eligible for Tier 2 services.

The possibility of the quadratic trend reflecting a qualitative difference in subgroups of children determined eligible for Tier 2 services is an interesting one and
perhaps lends support for the construct of learning disabilities. Qualitative information, obtained from the teachers, revealed consistent reports that the students’ level of cognitive ability was a primary factor in the perceived effectiveness of the Tier 2 intervention. Perhaps the slower rate of acquisition for those students performing near the right half of the quadratic curve can be explained in part by a pattern of cognitive weaknesses.

While the analysis yielded statistically significant findings for the effects of Tier 2 services when using the transformed data, the analysis using raw (untransformed) data did not. There are several factors that may have contributed to the disparity between these findings. The district’s use of the 10th percentile as their selection criteria may have inadvertently captured only those very students who would eventually require Tier 3 or special education services. Qualitative information, obtained from the teachers, revealed consistent reports that the students’ level of cognitive ability was one of the primary factors in the perceived effectiveness of the Tier 2 intervention. Lochman, Boxymeyer, Powell, Roth, and Windle (2006) indicate that “evaluating treatment is a complicated issue because poorer outcomes for some may be due to characteristics of the participants, such as low motivation or chaotic family conditions, rather than qualities of the intervention” (p 19). It is possible that the majority of these children would be classified as ‘non-responders’ and eventually provided more intensive instruction or perhaps, diagnosed with a cognitive or learning disability. Interestingly, a study using the 25% percentile as the selection criteria (Bryant et al, 2008) for Tier 2 intervention services in mathematics found non-significant effects for first grade Tier 2 services. This particular
study utilized a standard-protocol rather than a problem-solving approach for selecting the intervention; it is possible that the standard-protocol, ‘one-size-fits-all’ approach did not adequately target the individual needs of students. There is some support for this hypothesis, as research indicates that problem-solving method yields higher levels of achievement. Further, the duration of Tier 2 services may not have been adequate for these children to respond to the intervention. Several researchers (Vaughn, 2003; Vaughn & Roberts, 2007) suggest that 10 to 12 weeks is sufficient for determining progress in Tier 2; the duration of Tier 2 services for this study was 10 weeks, considered at the lower end in duration.

Another factor that may have contributed to the outcome of this study is that of floor effects. Floor effects occur when many individuals score near the lower end of the distribution, which can happen if a screening instrument is administered when children have not developed the cognitive maturity to perform on the assessment. This can lead to a high rate of over-identification and reduced accuracy in identification. A study conducted by Catts, Petscher, Schatschneider, Bridges, and Mendoza (2009) reveals that DIBELS measures (including ORF) were characterized by floor effects in the initial administration of measures. This finding is similar to research conducted by Silberglitt and Hintze (2005) who found floor effects for the ORF measure when it was administered in winter and spring of first grade. As this study utilized data from first grade student who were administered ORF probes in winter and spring, this may in part account for the non-significant findings for the analysis using raw data.
Utility of Regression Discontinuity

A second purpose of this study was to determine the utility of the regression discontinuity technique in studying the RTI model. There are several features of the regression discontinuity technique that may render it difficult to implement in educational settings. In regard to implementation, Trochim (1984) admonishes that:

The design is not easy to implement. While all research strategies are susceptible to poor implementation, in some ways the regression discontinuity design may be more sensitive than most. In this sense, the strength of the design may also be its greatest weakness. Its specific pattern of intervention into social reality must be followed strictly. The correct execution of the regression discontinuity design is threatened by social, political, and logistical problems in much the same way as in randomized experiments, and in some cases, the difficulties are more serious (p. 45).

One of the features of RTI is its flexibility, and some districts may move students in and out of Tiers as their performance on benchmark assessments dictates. This constant change of ‘sample’ may make it impossible to perform statistical analysis requiring intact groups. Further, school districts may view the strict adherence to the selection criteria in regression discontinuity as either a blessing or a curse; that is, the design may provide strict guidelines that help define the RTI model, or it may be too limiting to implement in practice.

Statistical analysis using the regression discontinuity technique rests largely on model specification. In regard to model specification, Trochim (1984) states:

There is no simple or mechanical way to determine definitively the appropriate model for the data. As a result, judgment and discretion on the part of the analyst are warranted. In addition, more experience in interpreting regression output
across a series of steps is needed...Prediction of the jump at the cutoff point is critical, but it is not desirable to accomplish this by generating complex models that make no substantive sense for the data (p. 148 – 149).

That there are no ‘clear cut’ rules for determining model specificity, may deter some researchers from using this technique. The regression discontinuity design requires the data analyst be familiar with regression and in some cases, the interpretation of non-linear data. Non-normal distributions may require data transformation, which renders interpretation difficult. Most teachers and some school psychologists may not have the training or background needed to analyze and interpret transformed or non-linear data.

Despite these limitations, the regression discontinuity technique has promise. Braden and Bryant (1990) discuss features of the design that render it a viable option for determining the nature of program effects for RTI data:

There are many features of regression discontinuity designs that school psychologists may find appealing, including: (a) the ability to test effects without withholding treatment; (b) data for the design may often reside in archives, group test scores, or other readily accessible media; (c) the design resists regression to the mean and other artifacts associated with treatment of individuals with extreme scores; (d) the ease of data analysis; and (e) the conceptual appeal of testing the effects of a program after selection criteria are taken into account (p. 7).

Many of the features of regression discontinuity (described above) are built into the RTI model. The 2006 article, New Directions in Research RTI (Response to Intervention): Rethinking Special Education for Students With Reading Difficulties (Yet Again) states in conclusion:
Although the regression discontinuity design seems counterintuitive at first, it has been endorsed by many of the leading research methodologists in the United States. RTI research provides a perfect venue for its use. We believe the logic behind it is sound, and it can be an excellent approach for experimental control studies of RTI that are easy to negotiate with schools (Section RTI and Regression-Discontinuity Designs: A Perfect Match (We Hope), ¶5).

Limitations of the Study

There are several limitations that impact the ability to generalize these results to other school districts using the RTI model.

1. The data used in this study consisted of archival data collected during the 2007-2008 school year. Qualitative information regarding fidelity of implementation was collected in a post-hoc manner.

2. This study was limited to first-grade students in one urban school district in southwestern Ohio. School programs “serve a finite group with a specific set of program resources” (Braden & Bryant, 1990, p. 7). Program effects may not generalize to other grade levels or a different demographic of student.

3. The outcome variable, spring administration of the ORF subtest, was the sole dependent variable. Other variables may produce different results.

Cognitive ability, a potential factor affecting the effectiveness of Tier 2 services, was brought to light by the intervention specialists and the school psychologist.

Cognitive ability may be an important covariate, but was not included in the regression model. Including this variable may have yielded different results.
Floor effects may be partially attributed to a lack of cognitive maturity, lack of attentional resources, or lack of linguistic skills (Catts et al, 2009).

4. This intervention in this study was implemented using a standard protocol approach; RTI models that incorporate a problem-solving model may yield different results.

5. This study used the 10\textsuperscript{th} percentile as the cut-off criteria for Tier 2 intervention services. Different cut-off criteria may yield different findings.

6. The study used data that had been transformed using logarithms, which renders interpretation challenging. As non-significant findings were obtained using the raw data, results should be interpreted with caution.

\textit{Suggestions for Further Research}

Based on the findings of this study, the following are suggestions for further research:

1. Future replications of this study may consider incorporating a measure of cognitive ability as a covariate when determining program effects.

2. A study comparing the program effects of a standard protocol approach and a problem-solving approach might assist school districts in deciding how to best implement the RTI model.

3. A study using a different assignment variable should be considered, as different variables may yield different results.
4. Important student demographics, including prior educational experiences (such as preschool or HeadStart), socio-economic status, and cultural variables should be included in future studies.

5. Future studies may consider implementing a ‘fuzzy regression discontinuity design’ if there is not strict adherence to the cut-off criteria.

6. Future studies may consider using a combination regression discontinuity/experimental design, where a random selection of participants scoring below the cut-off value serves as a control group.

7. A viable method which may reduce the model specification problem may include the use of double pre-tests for determining the pre-post functional form of the data (Trochim, 1984). Future studies may wish to include double pre-tests to assist with model specification. “If double pretesting is done, the model that best fits the pretest 1 – pretest 2 distribution could be directly applied to the pretest 2 – posttest analysis (Trochim, 1984, p. 151).

8. The number of students in the sample who later qualified for special needs services was not investigated in this study. Future studies may consider this when studying the effects of Tier 2 services.

Conclusions

The findings indicate that Tier 2 services appear effective in increasing achievement under the RTI model of service delivery. The results revealed a reduction in the percentage of students considered ‘at-risk’ on the second administration of the ORF
benchmark. Because RTI affects the LD category of special education most strongly, reduction in numbers of those children identified can be viewed in terms of a cost-saving measure to districts. For example, the outcome study conducted by VanDerHeyden et al. (2007), demonstrated that incorporating RTI practices resulted in a district-wide reduction in newly identified students from 6% to 3.5%, with associated placement costs reduced from $152,138.08 to $73,556.00. Increasingly, school districts have adopted the RTI model of service delivery as a method for providing early intervention and a method for identifying special needs students in an effort to remediate academic deficiencies before it becomes too difficult or costly. The nature of Tier 2 services should be further explored in effort to refine and develop the RTI model. The regression discontinuity design as an alternate method for determining the intervention effects in RTI should be further investigated and applied in various settings.
References


University of Oregon Center on Teaching and Learning. DIBELS (Dynamic Indicators of Basic Early Literacy Skills). Retrieved on 10/14/08 from http://www.dibels.uoregon.edu/dibelsinfo.php.


