ABSTRACT

PARKER, JUSTIN MICHAEL. The Use of Large-Scale Assessment Data to Investigate Special Education Effectiveness. (Under the direction of Ann Schulte).

This study used large-scale assessment data in reading and math for the state of North Carolina to investigate the effectiveness of special education programs. The study had two main goals. The first goal was to analyze how different methods of defining special education group status can impact magnitude of the achievement gap between special and regular education students. As first proposed by Ysseldyke and Bielinski (2002), it was predicted that the achievement gap would not grow over time in reading and math if group affiliation remained consistent across years. These hypotheses were confirmed. The second goal was to use a within subjects, ABA design to determine if students who transitioned in and out of special education in consecutive years performed better on the large scale assessments for the year they received special education services than the year before and the year after they received services. Results revealed that students performed better on the math assessment the year they received services than the year before they received services. However, no significant differences were identified for math scores for the year they received services compared to the year after they received services, and no differences were found for any of the reading comparisons. Results are discussed in terms of implications for reporting student progress and future research on special education effectiveness.
The Use of Large-Scale Assessment Data to Investigate Special Education Effectiveness

by
Justin M. Parker

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Psychology

Raleigh, North Carolina
2011

APPROVED BY:

__________________________
Ann Schulte, Ph.D.
Chair of Advisory Committee

__________________________
Susan Osborne, Ph.D.

__________________________
William Erchul, Ph.D.

__________________________
Jeffery Braden, Ph.D.
BIOGRAPHY

Justin Michael Parker was born in Oneonta, NY. He graduated from Edmeston Central School in 1997. In 2001, he graduated from the State University of New York, College at Cortland with a Bachelor of Science in Psychology (with honors). He earned a Masters degree in School Psychology from North Carolina State University in 2006. He is a member of the National Association of School Psychologists and of the North Carolina School Psychology Association.
ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Dr. Ann Schulte. I greatly appreciated the support, guidance and good humor throughout this process.

In addition, I would like to thank my committee members: Dr. William Erchul, Dr. Jeffery Braden, and Dr. Susan Osborne. I appreciated their willingness to provide me with constructive feedback and support when needed. And, I am very grateful for the help to make my study the best it could be.

Finally, I would like to thank my family, especially my wife, Dr. Alison Parker, and my parents, John and Sue Parker, for their unwavering support, love and encouragement through the years.
TABLE OF CONTENTS

List of Tables .................................................................................................................. vi
List of Figures .................................................................................................................. vii

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

Chapter Two ................................................................................................................... 4
  A Review of the Literature .............................................................................................. 4
    Special Education Legislation ......................................................................................... 4
      The Education for All Handicapped Children Act ...................................................... 4
      Compliance with PL 94-142: Special Education Identified as a Place ..................... 6
    Early Special Education Research ............................................................................... 7
    Classic Special Education Effectiveness Research ...................................................... 9
    The Regular Education Initiative and Inclusion ......................................................... 11
  A Call for Change ........................................................................................................... 13
  Cost of Special Education ............................................................................................. 14
  A Backlash Against Special Education ......................................................................... 15
  Current Strategies for Evaluating Special Education .................................................. 18
  What is Special Education? ............................................................................................ 19
    Operationally defining special education ................................................................... 20
    Using meta-analyses to group special education strategies ........................................ 22
  Defining Comparison Groups ....................................................................................... 25
  Identifying Group Composition ..................................................................................... 28

Chapter Three ............................................................................................................... 33
  Research Aims ............................................................................................................... 33
    Statement of the Problem ............................................................................................. 33
    Research Hypotheses ................................................................................................... 36
      Hypothesis 1 ............................................................................................................... 36
      Hypothesis 2 ............................................................................................................... 36
      Hypothesis 3 ............................................................................................................... 36
      Hypothesis 4 ............................................................................................................... 36
      Hypothesis 5 ............................................................................................................... 37
      Hypothesis 6 ............................................................................................................... 37

Chapter Four ................................................................................................................... 38
  Method ............................................................................................................................ 38
    Participants ................................................................................................................... 38
    Measures ....................................................................................................................... 42
      EOG-R ......................................................................................................................... 43
      EOG-M ......................................................................................................................... 44
    EOG test administration ............................................................................................ 46
Chapter Five

Results..................................................................................................................................47
Descriptive and Preliminary Analysis ...............................................................................47
  Merging Data Sets and Creating Cohorts ........................................................................47
  Creating Effect Sizes........................................................................................................48
  Normalizing Raw Test Scores and Creating Residualized Gain Scores..........................50
Analyses Specific to Hypotheses.......................................................................................53
  Comparing Methods of Reporting..................................................................................53
ABA Analyses .....................................................................................................................58
Summary of Hypothesis Testing .........................................................................................64

CHAPTER SIX.........................................................................................................................65
Discussion ............................................................................................................................65
Achievement Gap Comparison .........................................................................................66
  Summary of Findings (Hypotheses 1-4) ........................................................................66
  Discussion of Achievement Gap Findings ......................................................................68
    Different Trend Line for the Achievement Gap as Measured by the EOG-R ..............69
    Policy Implications for Subgroup Identification .........................................................70
Within Subjects, ABA Hypotheses ....................................................................................73
  Hypothesis Five ................................................................................................................73
  Hypothesis Six ...................................................................................................................75
General Discussion ...........................................................................................................79
  Addressing Limitations of Special Education Effectiveness Research .......................80
    Identification of Appropriate Outcome Measures ....................................................81
    Operational Definition of ‘Special Education Student’ .................................................83
  Limitations ........................................................................................................................85
    Research Design .........................................................................................................85
    Generalizability of Study Results ...............................................................................85
    Missing Information ....................................................................................................86
  Directions for Future Research .....................................................................................88
    Research Design .........................................................................................................88
    Similar Research Design with Different Operational Definitions .........................89
    Seek Alternative Measurements for Younger Students ............................................89
  Implications for Practice ...............................................................................................90
    Reporting of Student Progress ...............................................................................90
    Increased Access to Data for Decision Makers .......................................................91
    Realistic Expectations for Special Education Students .........................................92

REFERENCES .....................................................................................................................97
LIST OF TABLES

Table 1. Beginning Demographic Characteristics of the Sample for Hypotheses 1-4 .40

Table 2. Demographic Characteristics of the Sample for Hypotheses 5-6 (Students With Multiple Transitions Between Special Education and General Education) .................................................................42

Table 3. Means and Pooled Standard Deviations to Compute Effect Sizes ...............49

Table 4. Correlations Used to Compute Residual Gain Scores ................................51

Table 5. z-Score Means and Standard Deviations for Students Who Transitioned In and Out of Special Education by Cohort and Year ..........52

Table 6. Residual Gain Scores and Standard Deviations for Students Who Transitioned In and Out of Special Education by Cohort and Year ..........53

Table 7. The Effect Sizes and Confidence Intervals Across Grades Between General Education and Special Education Students for Students who were in Fourth Grade in 1998 ..................................................................55

Table 8. The Effect Sizes and Confidence Intervals Across Grades Between General Education and Special Education Students for Students who were in Fourth Grade in 1999 ..................................................................55

Table 9. F-values for Repeated Measures ANOVA by Cohort and Test Type ..........61

Table 10. Comparison of EOG Gain Scores Between the Year Students Were in Special Education (B) to the Year Before They Entered Special Education (A1) and to the Year After They Exited Special Education (A2) ...61

Table 11. Comparison of EOG Gain Scores Between Cohorts for the Year Before Students Received Special Education Services (‘A1’) ..................62

Table 12. Comparison of EOG Gain Scores Between Cohorts for the Year Students Received Special Education Services (‘B’) .................................................................63
## LIST OF FIGURES

Figure 1. Children 3 to 21 years old served by federally supported programs in the U.S. (in millions) ................................................................. 16

Figure 2. Students’ education status by year and grade level ............................ 41

Figure 3. Trend line comparison for EOG-R gain scores for students in the 1998 cohort .................................................................................. 56

Figure 4. Trend line comparison for EOG-R gain scores for students in the 1999 cohort .................................................................................. 57

Figure 5. Trend line comparison for EOG-M gain scores for students in the 1998 cohort .................................................................................. 57

Figure 6. Student Trend line comparison for EOG-M gain scores for students in the 1999 cohort ...................................................................... 58
CHAPTER ONE

Introduction

Over the past several decades, one of the most contentious issues among researchers and educators has been the effectiveness of special education (Dunn, 1968; Forness, 2001; Fuchs & Fuchs, 1994; Gartner & Lipsky, 1987; Leinhardt & Pallay, 1982). Special education was developed to improve outcomes for students with special needs who were barred access to education completely (McDonnell, McLaughlin, & Morrison, 1997) or were not benefiting from regular education (Stainback & Stainback, 1984). Although federal legislation assured the right to an education for students with disabilities in the 1970s, researchers have struggled to provide summative statements about the impact of special education on improving student outcomes. The current paper will provide a chronological overview of special education legislation, trends and research, followed by a proposal for a study that utilizes new data sources and innovative data-analytic strategies for examining outcomes for students with disabilities.

The review of the literature begins with a discussion of early special education legislation that was an extension of federal Civil Rights legislation. The right of children with disabilities to receive an education was assured in the 1970s by Public Law 94-142, the Education for All Handicapped Children Act (later called the Individuals with Disabilities Education Act). The purposes of the law and the rights that were guaranteed to children with disabilities will be presented, along with a discussion of how school districts’ initial focus on compliance with the law may have limited evaluations of student outcomes.

The review of the literature then examines classic special education effectiveness research. This section begins with a discussion of the methodological flaws characteristic of
Special Education Effectiveness

much of the special education effectiveness research, such as the inability to randomly assign children to treatment conditions and the difficulty in operationally defining special education. Next, results of early special education effectiveness research that suggested students with disabilities continued to struggle in the decade following the passage of PL 94-142 is discussed. The section concludes with a discussion of initiatives in the 1980s aimed at increasing the inclusion of students with disabilities in regular education classrooms.

The next section is an examination of the backlash against special education that occurred in the early 1990s. First, to set the stage for why the backlash against special education occurred, data that show the rising cost of special education are presented. Specifically, increased per-pupil expenditures and an increase in the number of students receiving special education services are presented as reasons for the increased cost of special education. The section concludes with a discussion of how the results of early special education effectiveness research and the increased cost of special education led to a backlash against special education in many popular press and scholarly articles.

The review concludes with an introduction of new strategies aimed at resolving some of the methodological concerns in special education research. The section begins with a discussion of how the increased accountability requirements for students with disabilities has resulted in unprecedented outcome data for special education effectiveness researchers. In this section, the strategies that researchers began to develop in order to address the methodological shortcomings that are common in special education effectiveness research are discussed. Specifically, strategies for operationally defining special education, identifying appropriate comparison groups, and methods for defining what students should be considered “special education students,” will be presented.
Following the review of the literature, Chapters Three, Four and Five describe the research aims and methodologies of the proposed study. We are currently in an era of increased accountability for student achievement. The use of large-scale assessments is a primary way that achievement data are collected for students. These assessments provide special education researchers with a wealth of outcome data for students with disabilities. Using large-scale assessment data from North Carolina, the current study will use new data analytic strategies in order to examine the effectiveness of special education in improving student outcomes. First, a method first proposed by Ysseldyke and Bielinski (2002) for defining group composition will be used in an investigation of the achievement gap between special and regular education students. Second, a within subjects design will be utilized to compare special education students’ achievement in the years they received special education services to the years they did not receive services. This study has the potential to provide evidence that the achievement gap between special and regular education does not increase over time as some have asserted (Shinn, 1986), and that special education policy has a positive impact on student outcomes.
CHAPTER TWO
A Review of the Literature

Special Education Legislation

Prior to the 1970s, many students with disabilities were not being provided with educational services (McDonnell et al., 1997). Through a series of Supreme Court rulings and Congressional legislation, students with disabilities were guaranteed access to a free public education that was not separate or different. The following section will briefly reviews Supreme Court rulings in civil rights and special education cases that led to Public Law 94-142, the Education for All Handicapped Children Act (1975). This section will also describe the original purposes of PL 94-142, and concludes with a discussion of how the focus on school districts’ compliance with the law led to special education being considered a place where services were received rather than a set of protections aimed at improving student outcomes.

The Education for All Handicapped Children Act

Historically, public education in the United States was the responsibility of state and local governments (Huefner, 2000). This began to change in the 1960s when, as a result of the Civil Rights movement and the Supreme Court ruling in Brown v. Board of Education (1954), the federal government began to offer financial help to the states in educating groups of students who were believed to be disadvantaged. The Civil Rights movement also served as a catalyst for efforts to offer a public education to students with disabilities. Disability advocates argued that students with disabilities should be afforded the same protections as minority students in terms of access to education (Huefner, 2000). In 1972, two Supreme Court cases, Pennsylvania Association for Retarded Children (PARC) v. Pennsylvania (1972)
and *Mills v. District of Columbia* (1972), ruled that school districts were required to provide access to education for students with disabilities. In response to these landmark rulings, Congress passed legislation designed to assure that children with physical, cognitive, and behavioral disabilities were not excluded from public education. Public Law 94-142, the Education for All Handicapped Children Act (1975), established that all children with disabilities were entitled to a free and appropriate education. Although the law established that all children with disabilities had the right to receive an education, Congress acknowledged the difficulty in making a global statement on how to best serve and educate children with disabilities. Therefore, decisions about the placement settings and educational programs for children with disabilities were to be made on an individual basis by a team of parents and professionals (McDonnell et al., 1997).

The original Education of Children with Disabilities Act (later called the Individuals with Disabilities Education Act, IDEA; 1975) had four major purposes: (a) to assure that all children with disabilities have available to them a free public education that emphasizes special education to meet their unique needs; (b) to assure that the due process rights of children with disabilities and their parents are protected, including the right of notice and consent prior to actions affecting their child, and the right to a due process hearing to resolve complaints and disagreements between parents and the school; (c) to provide financial assistance to states and localities to provide the education of all children with disabilities, and; (d) to assess and assure the effectiveness of efforts to educate all children with disabilities. Although there have been several reauthorizations of the original PL 94-142 (e.g., IDEIA, 2004), the basic purposes of the law have not been altered (Smith, 2005).
The passage of PL 94-142 guaranteed that students with disabilities would receive educational services. However, the ambiguous language of the law helped fuel further debate about how children with disabilities were to be served. For example, how to balance the dual mandates of an *appropriate* education and an education that takes place in the *least restrictive environment* has been a focus of considerable debate in education (Scruggs & Mastropieri, 1996). Similarly, although the legislation stated that the federal government would pay up to 40% of the costs related to educating students with disabilities, an exact dollar amount was not provided (Smith, 2005). Finally, and perhaps most important to special education effectiveness research, the law did not state that the government would assess special education effectiveness in terms of student outcomes, but rather the effectiveness of *efforts* to educate children with disabilities. Therefore, many of the early special education evaluations were focused on compliance with federal legislation rather than special education effectiveness (Borich & Nance, 1987).

*Compliance with PL 94-142: Special Education Identified as a Place*

In the initial stages of implementation of PL 94-142, the focus of much of school districts’ activity related to special education was establishing the procedures for implementing the law and outreach to assure that all children with disabilities were identified and served (Smith, 2005). Although the procedural focus allowed schools to show evidence of their compliance with federal legislation, it did not allow for an examination of the effectiveness of special education. That is, schools could show evidence that more students had received special education services as a result of the new procedures, but there was no mechanism to show that special education was improving the educational outcomes of students with disabilities (Mastropieri & Scruggs, 2000).
Given the procedural focus of special education evaluation, special education was increasingly seen as a place to provide students with disabilities services rather than a policy that intended to improve the academic achievement of students with disabilities (Schulte, Osborne, & Erchul, 1998). This view began to change in the early 1980s, fueled by research that indicated children with disabilities were still struggling academically nearly a decade after PL 94-142. Much of the research during this period focused on a comparison of the different settings where children with disabilities received their services. A distinction between a separate settings placement and a general education placement was typically made by considering the percentage of the school day a child spent in the general education classroom (Mastropieri & Scruggs, 2000), although there was no set criterion for this percentage between research studies (Carlberg & Kavale, 1980). Gradually, researchers began to realize that the place where services were received was less important than the services themselves, and therefore “place” was not a very meaningful variable in special education effectiveness research (Leinhardt & Pallay, 1982). The next sections will review the recent history of special education following the passage of PL 94-142. This review will include classic special education research, the backlash against special education that was in part due to the methodological shortcomings of this classic research, and end with current data-analytic techniques aimed at improving special education effectiveness research.

Early Special Education Research

With the ushering in of IDEA, many students with disabilities who once were barred access to educational services were now being educated in public school systems. Given that one of the reasons for the development of special education was to improve student outcomes, and a goal of IDEA was to evaluate the effectiveness of efforts to educate students
with disabilities, researchers became increasingly interested in examining the effectiveness of special education programs in terms of student outcomes (Forness & Kavale, 1994; Fuchs & Fuchs, 2005; Madden & Slavin, 1983). Before reviewing the classic special education research, it is important to consider the methodological flaws that limit the interpretability of this research.

First, many special education effectiveness research studies had small sample sizes, which limited the ability to interpret and generalize results. Second, many studies compared academic outcomes for students without disabilities in general education to the academic outcomes of students in a variety of special education settings. Given that students are often identified as having a disability because they have difficulty meeting academic and social demands more than students without disabilities, comparing the outcomes of these two groups may not be appropriate (Ysseldyke & Bielinski, 2002). Third, few of the studies employed random assignment of children to the service delivery settings. Although a methodological shortcoming, ethical considerations make this limitation difficult to overcome. IDEA stipulates that each child must be afforded the opportunity to succeed by getting the services they need, and thus random assignment in special education effectiveness research seems impossible. Fourth, much of the research failed to adequately describe the instruction provided to the students in the control and treatment groups (Leinhardt & Pallay, 1982), and did not adequately define the disability status and characteristics of the groups (Forness & Kavale, 1994).

Unfortunately, these methodological shortcomings severely limit the conclusions that can be drawn from this body of research. Furthermore, a lack of a universal definition of special education (a limitation that will be discussed in greater detail later in this review)
makes it difficult to understand what comprised the independent variable *special education* when reviewing the research literature. Of the methodologically sound studies, results indicated that (a) academic outcomes for students with disabilities were not positive regardless of where instruction took place (Schulte et al., 1998); (b) students with disabilities in inclusive classrooms tended to have better academic and social outcomes than did students who received services in more restrictive settings, although this finding was inconsistent (Carlberg & Kavale, 1980; Madden & Slavin, 1983); and (c) the type of instruction given within a setting was a better predictor of academic and social outcomes for students with disabilities than was the placement setting itself (Leinhardt & Pallay, 1982).

*Classic Special Education Effectiveness Research*

One of the earliest and most influential studies that examined special education effectiveness was a meta-analysis conducted by Carlberg and Kavale (1980). The authors sought to compare outcomes for students with disabilities who received their special education services in general education classrooms (referred to as mainstreamed or inclusive classrooms) to outcomes for students who received their educational services outside of the general education classroom. Their meta-analysis revealed that placement in a separate setting was inferior in terms of academic outcomes for children with disabilities to placement in a general education classroom (overall effect size of -.12). However, the outcomes varied by disability type. Children with low cognitive functioning (slow learners and educable mentally retarded students) performed better in inclusive classrooms (the effect size for special class placement for these groups was -.34 and -.14 respectively), but children with learning disabilities and behavioral/emotional disorders had better outcomes when they received special resource room services (the effect size for special class placement for these
students was +.29). However, because of the small number of studies that considered
disability type in this meta-analysis, the authors urged caution when interpreting the results
of this study by disability classification. Although many of the studies that were part of the
meta-analysis were methodologically flawed, the results of this study raised concerns about
whether or not special education was a useful tool for helping students.

Another classic publication in the special education effectiveness literature was a
review by Madden and Slavin (1983). Noting the methodological shortcomings of much of
the research literature regarding academic outcomes for students with disabilities described
above, the authors conducted a critical review of studies that compared the effects of placing
students with mild academic disabilities in separate settings versus full-time inclusive (what
they called mainstreamed) classrooms. Relying primarily on two methodologically strong
studies, where differences in teacher skill and curriculum had been controlled, the authors
concluded that general education settings produced better academic outcomes for students
with disabilities than separate settings. However, in these studies, instruction had been
adapted to meet children’s needs in both types of settings. In other studies, where instruction
was not adapted or extreme differences between settings were not controlled, the results were
less clear. Thus, an important aspect of the observed positive outcomes for students with
disabilities in general education classrooms in the Madden and Slavin (1983) review was
appropriate instruction.

conducted a meta-analysis that examined academic outcomes of students with a range of
disabilities receiving services across a variety of settings. They found that students in
inclusive classrooms made greater academic gains in terms of academic achievement than
students with disabilities in either resource room or separate settings classrooms. Similarly, Leinhardt and Pallay (1982) examined the effectiveness of different special education settings for children with a wide range of disabilities. In their review, the authors suggested that setting is not a particularly meaningful variable in understanding academic outcomes for students. They concluded that positive academic outcomes can occur regardless of the educational placement. They asserted that it was the social and instructional processes that occurred within a setting that determined the educational outcomes for students with disabilities.

*The Regular Education Initiative and Inclusion*

From a global perspective regarding academic outcomes of students with disabilities, a growing body of research indicated that children with disabilities were at risk for poor academic outcomes regardless of their placement setting (Madden & Slavin, 1982; Zigmond et al., 1995; Schulte et al., 1998). Given the lack of evidence that showed segregated special education placements were superior to general education placements, some researchers and educators began to push for more inclusive placements for students with disabilities. In what became known as the Regular Education Initiative (REI), educators advocated for a merging of special and regular education in order to serve children with mild disabilities in regular education classrooms (Will, 1986). With the REI, students with mild disabilities, who were the majority of students in special education, would no longer receive their special education services in separate settings. The goal of the REI was to give children the advantage of being in a general education classroom while still receiving the supports needed in order to succeed in both the academic and social realms. Thus, the needs of students’ with mild disabilities could be met in the general education classroom on an
individualized basis, with regular education teachers and special education teachers working together to serve students (Gartner & Lipsky, 1987).

Others during this time argued that all students with disabilities should be in the regular education classroom at all times. Proponents of “full inclusion,” a more radical movement than the REI, argued that removing children from the regular education classroom, even those with severe disabilities, conveyed the message that children with disabilities were different, and therefore it was acceptable to segregate them from other students. To many supporters of full inclusion, removing a child from the general education classroom was morally wrong (Fuchs & Fuchs, 1998).

Despite the argument from advocates of the full inclusion movement that students should not receive any of their services in separate settings, full inclusion received limited empirical support as a mechanism for improving academic outcomes of students with disabilities. Research indicated that general education teachers were unlikely to adjust their instructional styles in order to address the wide range of disabilities that children may present in their class (Baker & Zigmond, 1990; Fuchs, Roberts, Fuchs, & Bowers, 1996). Without instructional adjustments by teachers, students with disabilities were unlikely to receive the support they needed in order to achieve academic success in general education classrooms (Leinhardt & Pallay, 1982). Furthermore, even when general education teachers used best practices in inclusive classrooms, research suggested that many children with disabilities still needed intense and direct intervention that was best delivered outside the general education classroom (Fuchs & Fuchs, 1998). As a result of evidence that suggested a full inclusion model did not provide students with disabilities the support they needed, the movement failed to gain momentum the way that the REI did. Although some researchers continued to support
a full inclusion model, most educators and researchers supported a less radical approach to including students with disabilities in regular education classrooms.

In sum, there is a wealth of research that compares academic outcomes for students with disabilities as a function of the setting in which they were educated (Fuchs & Fuchs, 1994). Although this type of comparison was important and led to the REI, it does not answer the central question of special education effectiveness. Special education was created to be a set of protections and specialized services that are designed to improve outcomes for students with disabilities. When evaluating special education effectiveness, researchers need to consider special education as a policy rather than a place where students go to receive services. As researchers continued to grapple with ways to evaluate the effectiveness of special education in the 1980s and 1990s, the number of students who received special education services continued to grow. As a result, the cost of special education continued to grow as well. During this time, the rising cost of special education combined with data that indicated students with disabilities continued to have poor academic outcomes resulted in a public backlash against special education.

A Call for Change

As discussed in the previous section, classic research in special education typically focused on where students with disabilities received their special education services. Although a review of this literature reveals mixed results about what type of placement setting is superior, and the research was typically methodologically flawed, a consistent finding across most studies was that students with disabilities performed at a level that was below that of their non-disabled peers (Zigmond et al., 1995). Although many studies found that the academic achievement of students with disabilities could be improved when proper
supports were in place (Leinhardt & Pallay, 1982), there still existed an achievement gap between students with and without disabilities. In fact, this achievement gap often grew larger as students progressed to upper grades (Shinn, 1986). Taken together, the research indicated that outcomes for students with disabilities were not positive (Schulte, 1996).

In the early 1990s, these findings along with a renewed interest in the cost of special education combined to create a backlash against special education. Popular press writings fueled this backlash by writing scathing articles about the ineffectiveness of special education despite the escalating costs. In addition, many scholarly articles, often citing methodologically flawed research, argued that special education was not only ineffective, but that it harmed children with disabilities and was therefore morally wrong (Gartner & Lipsky, 1987; Skrtic, 1991; Sleeter, 1986). The backlash was so strong that it led the National Association of State Boards of Education (1992) to declare, “It is now time to ask, are children currently classified as special education students achieving what they are capable of? Are they being prepared for life after school? Are current mainstreaming practices producing their intended outcomes? We answer, no!”

This section will provide a review of the rising cost of special education in order to frame the backlash against special education that occurred in the early 1990s. The section will conclude with a brief review of the popular press and scholarly articles that suggested special education was ineffective, not cost effective and, in some cases, immoral.

Cost of Special Education

In the early 1990s school expenditures were four times what they had been less than four decades earlier, even when inflation was controlled for (Guthrie, 1997). However, as other school programs fell victim to budget cuts, special education funding continued to
Special Education Effectiveness

exceed the costs of general education (Fuchs & Fuchs, 2005). The cost of educating a student in special education was more than double the cost of educating a student in general education (Chaikind, Danielson, & Brauen, 1993), a ratio that had remained relatively constant since the passage of IDEA in the 1970s (“Special Ed’s Special Costs,” 1993; Chambers, Parrish, & Harr, 2002). In the 1989-1990 academic year, the national average per-pupil cost of special education and related services was $7,800, or about 2.3 times the cost of regular education (Fuchs & Fuchs, 2005).

Exacerbating the extra per-pupil cost of special education was the fact that the number of students served by special education services continued to grow. According to the National Center for Education Statistics (2006), 3.7 million students received special education services in the United States in the 1976-1977 academic year. By the 1993-1994 academic year, the number of children who received special education services increased to approximately 5.2 million students (Figure 1; more recent data are included to show the continued trend). The data revealed that students with learning disabilities made up a large proportion of the increase in the number of students being served by special education. The large proportion of students with learning disabilities in the general education population is an important consideration, as the next section will discuss some researchers’ belief that many disabilities, especially mild disabilities, are not disabilities at all but are instead social constructs created for societal reasons (Skrtic, 1991).

A Backlash Against Special Education

Two decades after the original IDEA was passed into law, a growing body of research suggested that academic outcomes for students with disabilities were still not positive. Combined with the rising cost to educate students with disabilities, this research fueled a
backlash against special education. Popular press magazines, scholarly articles and general education groups began to attack special education and label it as an ineffective policy. In a particularly critical article, *U.S. News and World Report* (Schapiro et al., 1993) stated that special education programs are often designed to attract federal dollars, not to best serve students with disabilities. The article went on to say that in special education classrooms, academics often take a back seat.

*Figure 1.* Children 3 to 21 years old served by federally supported programs in the U.S. (in millions).

This backlash against special education was not limited to popular press. Many scholarly articles, including Gartner and Lipsky (1987), questioned the legitimacy of special education and accused special education policy of treating students with disabilities unfairly (Hallahan & Kauffman, 1994). Some researchers suggested that special education should be
eliminated completely because research continued to show that it did not improve student achievement (Stainback & Stainback, 1992). Others argued that many disability classifications, and thus special education classifications, were not true disabilities but rather societal constructs (Skrtic, 1991). Sleeter (1986) argued that learning disabilities were created as a way of explaining the failure of some students to succeed in light of increased academic standards:

Students unable to keep up with raised standards were placed into one of five categories. Four were used to explain the failures of lower class and minority children; learning disabilities was created to explain the failures of white middle class children” (p. 46).

The popular press and scholarly writings that attacked special education served to feed a zeitgeist that special education was not only ineffective, but also harmful and unethical (Fuchs & Fuchs, 2005; Hallahan & Kauffman, 1994). As a result, researchers in special education began to look for new ways to demonstrate that special education was indeed effective. Since the passage of IDEA (1975), special education at federal, state, and local levels had been primarily concerned about compliance with the Free and Appropriate Public Education dimension of IDEA, largely ignoring student outcomes (Hehir, 1994). The public outcry against special education led many in the field to acknowledge that too little attention had been paid to student outcomes. Although many researchers did not agree with the assertion that special education was not and could not be effective, they found it difficult to convey this message to educators and the general public. This view was summarized by Fuchs and Fuchs (2005), who said:
(People) must hear from more than special education bashers; those who believe the field is worth preserving must respond vigorously to persons outside and inside the profession who would have special education dismantled and hauled off to the town dump. And each of us must assume this responsibility personally because our so-called leaders appear to have a terminal case of the shakes (p. 364).

Accordingly, many special education researchers began to focus on new ways to investigate special education effectiveness.

Current Strategies for Evaluating Special Education

A national educational trend toward increased accountability has led many states to implement large-scale assessments as a way to track student progress (Schulte & Villwock, 2004). These assessments have the potential to provide researchers with valuable outcome data for students as they progress through the grade levels. Given concerns about the effectiveness and rising costs of special education, these data were particularly enticing to special education researchers. However, students with disabilities were often excluded from the assessments, and most discussions about special education continued to be centered on cost rather than effectiveness (Vanderwood, McGrew, & Ysseldyke, 1998). As such, gathering data to examine special education effectiveness was not a paramount concern of legislators (Hanushek, Kain, & Rivkin, 2002). This began to change in 1997 when the reauthorization of IDEA required students to participate in state- or district-level assessments (Schulte & Villwock, 2004). In 2001, the No Child Left Behind Act required that states disaggregate data on the performance of students with disabilities with the same detail and frequency as they report the performance of students without disabilities (Ysseldyke & Bielinski, 2002).
Because students with disabilities are now required to be included in large-scale assessments, researchers in special education have a new tool to use to investigate student outcomes. However, the problem facing special education researchers is how to best use these data, as having access to the data does not solve the methodological shortcomings that have plagued special education research. Furthermore, it seems evident that there is not one simple answer to address the methodological problems facing the special education effectiveness research. Rather, varied approaches should be used to address specific limitations, with each approach contributing a piece of the puzzle. The next sections will discuss strategies for overcoming three methodological shortcomings still facing special education. Following a discussion of a method for operationally defining special education, the final two sections of the literature review will discuss strategies for selecting appropriate comparison groups and identifying group membership when using large-scale assessment data in special education effectiveness research.

*What is Special Education?*

At the most basic level, researchers must be able to operationally define the variables they are investigating. As a policy, special education can be defined as a set of protections designed to improve the academic performance of students with disabilities. If a traditional research design were used to study special education effectiveness, students with similar abilities would be randomly assigned to either receive special education services or receive no special education services. The outcomes of these students would then be compared and a statement could be made regarding the effectiveness of special education. However, given that IDEA requires an appropriate education for all students, it is illegal (and many would say unethical) to deny a child special education services, even to study the effectiveness of the
services the child was to receive. Furthermore, even if special education researchers were able to use randomized trials, there is little consensus as to what should be compared. That is, when evaluating the effectiveness of special education, it is difficult to determine what makes up the “set of protections” for students with disabilities.

Unfortunately, operationally defining special education in effectiveness trials is a difficult task. One strategy for defining special education is to break the policy down into smaller parts. For example, some researchers have investigated the effectiveness of individual interventions that are often implemented as special education services (e.g. Forness, 2001). This strategy allows for a comparison of strategies within special education, thus providing evidence of what strategies are most effective to educators and researchers. However, breaking down special education into smaller pieces does not allow for global statements about the effectiveness of special education. Given that recent educational trends of accountability have included students with disabilities (NCLB, 2001), more global statements about the effectiveness of special education are necessary as well (Parker & Hagan-Burke, 2007). However, investigations of special education effectiveness as a policy typically do not provide detailed information about what strategies are most effective, and are therefore of limited use to educators. Thus, researchers must use both strategies in special education effectiveness research. One strategy for operationally defining special education by breaking it down into smaller pieces is presented next, followed by a discussion of strategies for overcoming methodological flaws in more global special education effectiveness research.

*Operationally defining special education.* When considering the varied findings across research studies, it is important to remember that the term “special education” does
not have a universal meaning, and therefore different studies that refer to special education may not be talking about the same construct. That is, the lack of a universal definition and the many ways that special education is implemented make it difficult to generalize the findings of individual research studies. As discussed earlier, in addition to investigating the overall effectiveness of special education, studies have compared outcomes of students with disabilities based on their placement setting, disability type, level of instruction and type of intervention used. Although some classic research can be generalized to the field of special education, such as Leinhardt and Pallay’s (1982) finding that placement setting was not a useful construct, most studies are much more limited in scope because of the number of variables that cannot be controlled. For example, when comparing the effectiveness of one intervention to another, researchers are unable to control for teacher variables, school variables, and differences in abilities between the students. Even comprehensive reviews of the literature make it difficult to identify whether an intervention is more effective for certain types of children, is better for certain types of problems, or is more effective than other interventions because it is difficult to generalize the findings from one setting to another (Forness & Kavale, 1994).

In an attempt to parcel out the ways in which researchers have investigated special education effectiveness, Forness (2001) focused on the type of interventions students with disabilities received. He argued that there are three ways in which special education interventions have been defined in the research literature. These definitions were based on whether “special,” “education,” or “related” was emphasized when describing special education and related services. Research that seemed to emphasize “special” investigated the effectiveness of interventions that would not be routinely used in general education. That is,
they were developed to enhance unobservable constructs that were thought to be the cause of learning deficits. The goal of these interventions was to improve skills that underlie academic learning (e.g. learning strategies), not specific knowledge acquisition. Interventions that focused on “education” were those that modified and adapted general education instruction to meet the needs of children in regular education classrooms. These interventions relied on direct instruction with the goal of improving the academic performance of special education students. Interventions that focused on “related” were dependent on interventions not delivered by teachers. Children who received this type of intervention were often out of the regular education classroom in order to receive services. A major category not included in his overview was the role of “place” in special education. As can be seen from the previous review of the literature, place is not a particularly useful variable when investigating special education effectiveness. Although many of the interventions in the “related” category were delivered to children outside the regular education classroom, the focus of the research was on the interventions themselves rather than the place they were received.

*Using meta-analyses to group special education strategies.* By grouping special education effectiveness research according to the type of intervention that students received, Forness (2001) provided a good overview of how special education has been defined and investigated. As discussed earlier, special education services vary greatly across settings, making it is difficult to generalize the findings of a single-site special education effectiveness study. Furthermore, these variations make it difficult to determine if one intervention is superior to another when reviewing the special education effectiveness literature. Therefore, meta-analyses are a useful tool for synthesizing the research findings. They allow researchers to combine the findings of several studies, thus allowing for a comparison of intervention
types that, because of the larger sample size, is not as susceptible to variations across the research settings. By combining the findings of many meta-analyses into a “mega-analysis,” special education researchers are able to look at the effectiveness of special education in terms of effect sizes and compare the effectiveness between interventions, disability type, and setting.

Forness (2001) utilized 24 meta-analyses in order to draw conclusions about the effectiveness of special education interventions. Before discussing the results of this study in detail, it is important to consider how the author combined results from single-case research (SCR) with results from group research. In his review, Forness used a regression approach to determine the effect sizes in each study. Unfortunately, SCR data typically do not meet the parametric data assumptions of normality, equal variance and independence (Parker, Hagan-Burke, & Vannest, 2007). Therefore, a regression analysis is inappropriate for combining results from SCR with group research. Forness (2001), acknowledging that combining the results of SCR with group research was a controversial topic, urged caution when interpreting the results of his “mega-analysis.” “Not only are there hazards in combining such diverse meta-analyses, but such a process also ignores the variegated picture of special education efficacy” (p. 191). Despite this limitation, the method that the author used to separate out pieces of special education services provides a good model for how researchers can operationally define special education, and therefore a discussion of the results of his study is warranted here.

In his mega-analyses, Forness (2001) found an overall special education effect size of .55. The effect sizes of the individual interventions that were included in the analysis varied greatly, from 1.62 (mnemonic strategies) to -.12 (special class placement). The author
argued that this variation in effectiveness between interventions contributes to the critics’ point of view that special education is ineffective, as certain interventions have been shown to be ineffective. In order to illustrate that it is not special education in general that is ineffective, the author grouped the interventions by effect size. Of the 24 interventions types, eight had a “small” effect size (less than .4), eight had a “medium” effect size (.4 to .79), and eight had a large effect size (greater than .8).

Despite the overall effect size of .55, Forness (2001) noted that the results of the mega-analysis may be misleading due to the many strategies that are used under the umbrella term special education. It was therefore important to identify what part of special education the interventions focused on (“special,” “education,” or “related,” as discussed earlier). When grouping the interventions in this way, the author found the “education” strategies had an overall effect size of .84, the “related” strategies had an overall effect size of .53, and the “special” strategies had an overall effect size of .20. By grouping the intervention types in this way, the author was able to operationally define special education according to what type of service the students were receiving. With this approach, Forness was able to identify what strategies were effective within special education, therefore challenging critics’ global statements that special education was an ineffective policy. He argued that special education was not an ineffective policy, but rather there were specific interventions within special education that were less effective than others.

In sum, despite methodological concerns, the results of this mega-analysis revealed the importance of defining what was meant by special education. Forness (2001) acknowledged that special education included so many variables that it was very difficult to make global statements about the effectiveness of special education as a whole. Instead, the
Special Education Effectiveness

author chose to investigate a piece of special education (intervention type), and thereby developed a model of how to operationally define a portion of special education.

Although mega-analyses can provide important data, this method of examining special education effectiveness does not address methodological concerns in the actual research. That is, mega-analyses are useful for interpreting the research, but the results of the mega-analyses may still be limited by methodological shortcomings in the original research. Two of the most prevalent methodological shortcomings in special education effectiveness research continue to be the difficulty with finding an appropriate comparison group and group selection strategies. The following sections will discuss strategies for overcoming these shortcomings in more global special education effectiveness research.

Defining Comparison Groups

A major obstacle in special education effectiveness research is the difficulty of isolating causal effects of special education (Hanushek et al., 2002). By definition, special education and regular education students differ in significant respects, making a comparison of these groups not a valid measure of special education effectiveness (Blackorby & Wagner, 1996). Although federal legislation mandates that students with disabilities have academic goals that are closely aligned with goals set forth for students without disabilities, research suggests that students with disabilities struggle more academically than students without disabilities. Therefore, closing the achievement gap between special and regular education students is difficult (Shinn, 1986). For example, Ysseldyke et al. (2004) reported that a substantially smaller percentage of special education students meet state standards as compared to general education students. Thus, simply comparing outcomes of those two groups to answer the question of special education effectiveness is not appropriate. In order
to understand how children with disabilities are performing as a subgroup, it is important that their scores are disaggregated from students without disabilities (Ysseldyke & Bielinski, 2002). Because students with disabilities are now included in most large-scale assessments of achievement (Schulte & Villwock, 2004), researchers have access to unprecedented outcome data for this group of students.

One of the first studies to use large-scale assessment data to investigate special education effectiveness was conducted by Hanushek, Kain, and Rivkin (1998). In what has become an influential study on how large-scale assessment data can be utilized, the authors sought to identify an appropriate comparison group for special education students. They argued that following students who move in and out of special education programs could allow researchers to control many of the extraneous variables between groups that often limit special education effectiveness research. In their study, the authors used end of grade test data to compare the academic growth of students in years they were in special education to their growth in years they did not receive special education services.

It should be noted that in addition to examining the effects of special education services on all identified students with disabilities, Hanushek et al. (1998, 2002) also investigated the special education program effects of students with learning, speech/language, and behavioral disabilities separately. The authors provide three reasons why these categories were chosen to be examined separately. First, these students accounted for the largest disability categories. Other disability categories had small sample sizes and very few students who transitioned in and out of special education. Second, the chosen disability categories were those that schools had the most discretion over. Therefore, any decisions to expand or contract special education would specifically target these disability
categories. Third, the authors expected to have a much greater impact on students with learning disabilities than on students with speech problems. By including both groups of students, the authors hoped to show that special education was able to improve academic outcomes for students where academic improvement was their primary goal, while the academic outcomes for other types of students in special education would be less affected.

Results of their study revealed that special education increased students’ end of grade state test scores in math and reading (Hanushek et al., 1998). When estimating the special education program effects for all special education students, the authors report that one year of special education programming improved student performance by .1 standard deviations in math and by at least .04 standard deviations in reading. That is, at the end of one year of special education services, students’ who received special education services closed one-tenth of the achievement gap in math and nearly one-twentieth of the gap in reading. When the authors compared the end of grade test scores of students from the year they entered special education to the year they exited special education, results revealed that the math test scores for entrants increased .17 standard deviations, and the scores for students who exited increased just .02 standard deviations. Results from the reading tests were similar, although less robust. Reading test scores for students who entered special education increased .08 standard deviations, and test scores of students who exited special education increased .01 standard deviations.

When the authors disaggregated students according to disability type and compared scores for the year of entry to the year of exit from special education, they found that students with learning disabilities had the biggest increase in end of grade math test scores. The math test scores of students with learning disabilities who entered special education increased .18
standard deviations, and test scores of those who exited increased .06 standard deviations. Math test scores of students with speech impairments who entered special education that year increased just .02 standard deviations, and test scores of those who exited special education increased .01 standard deviations. These results confirmed the authors’ hypothesis that special education would have its largest achievement impact on students with learning disabilities and a much smaller impact on students classified as speech impaired. However, there were no significant differences between the groups in reading test scores.

In summary, Hanushek et al. (1998, 2002) addressed the difficulty of finding appropriate comparison groups in special education research by tracking the progress of students who transition in and out of special education. Rather than using regular education students as a control group, the authors used large-scale assessment data to investigate special education effectiveness by comparing the achievement gains of students while receiving special education services to gains of students while not receiving special education services. They found that students with disabilities benefited from receiving special education services, especially in terms of their test scores in mathematics.

**Identifying Group Composition**

As shown by Hanushek et al. (1998, 2002), students who move in and out of special education can allow researchers to have an appropriate comparison group in special education effectiveness research. However, the transient nature of special education makes it difficult to define what students should be included in the special education group in a given year. Therefore, it is difficult to interpret program effectiveness by following individual students across academic years. This problem is compounded when researchers attempt to
compare groups of students, as students’ membership in a group (special education or regular education) may change from year to year.

Ysseldyke and Bielinski (2002) proposed two strategies for defining special education group membership for research that follows student achievement across multiple years: the cohort-static and cohort-dynamic approaches. In the cohort-dynamic approach, all students in a particular year who received special education represent the special education group. That is, students’ group membership is redefined each year according to whether or not they received special education services that particular year. In the cohort-static approach, special education membership is defined by a student’s special education status in the first year of the study’s testing, and does not change from year to year. Therefore, if a student received special education services in the first year of the study, that student would be considered a special education student for the remainder of the study, regardless of whether or not he or she received services after the first year of the study.

The authors argued that student transitions in and out of special education could make the achievement gap between regular and special education students appear larger than it truly is. The highest achieving special education students are more likely to exit special education than lower achieving students. Because the cohort-dynamic approach redefines group membership each year, the highest achieving special education students who exit special education would not represent the special education group in subsequent years. Therefore, even though special education may have helped a student in terms of academic achievement, the gain made in achievement would not be reflected in the special education group’s data.
Given that the highest achieving students are the most likely to exit special education, Ysseldyke and Bielinski (2002) sought to demonstrate how transitioning students can affect results in achievement gap research. In their study, the authors compared the cohort-static and cohort-dynamic approaches when evaluating trends in the achievement gap using statewide achievement data from the Texas Assessment of Academic Skills (TAAS). The TAAS assesses reading, writing, mathematics, social studies and science, and results are reported on the Texas Learning Index (TSI), which is similar to a grade level $z$-score. That is, the TSI translates the original raw score distributions for each grade into scaled scores with a fixed mean and standard deviation.

Ysseldyke and Bielinski (2002) compared mean TAAS scores of special education students as identified using the cohort-dynamic approach to scores of special education students as identified using the cohort-static approach. Mean TAAS scores of general education students were used as the reference group in order to determine the effect size for special education services. Results revealed that when the cohort-static method was used, the special education effect size as measured by the reading test stayed relatively stable from 3rd to 8th grade. The mean score for special education students was approximately one-half standard deviation below the mean for general education students in both third (-.48) and eighth (-.42) grades. When the cohort-dynamic approach was used and the same data were evaluated, the mean score for special education students dropped from one-half standard deviation (-.48) to almost a full standard deviation (-.93) below the mean score of regular education students. Although the authors indicated that roughly 80% of the special education population remained unchanged in any given year, it was important to include students who exited special education as members of the special education group when tracking student
progress over multiple years. Furthermore, it is important to recognize that students exiting special education outperformed those entering special education by .75 standard deviations. Thus, when the authors used the cohort-dynamic approach, the special education group became a more concentrated group of low-achieving students.

Although Ysseldyke and Bielinski (2002) argued that a cohort-static approach was superior to a cohort-dynamic approach when investigating the achievement gap between regular and special education students, they noted that the cohort-static approach might not be satisfactory in order to make statements about the effectiveness of special education. In this approach, it is impossible to determine whether or not the achievement gains should be credited to special education services or to general education services. They suggested that states’ large-scale assessment data might contain large numbers of students who transition between special education and regular education more than once. These transitions should make it possible to use an ABA design to measure special education effectiveness, thus expanding on the design of Hanushek et al. (1998, 2002). This type of design would allow for: (a) students to serve as their own controls, thus avoiding the problem of finding a valid comparison group; (b) data to show when gains were made, thus allowing the researcher to attribute gains to special education services or regular education services; and (c) analysis of a high-achieving subgroup of special education students (relative to the achievement of all special education students), as the high achieving students tend to transition more frequently in and out of special education.

In sum, Hanushek et al. (1998, 2002) and Ysseldyke and Bielinski (2002) provided researchers with new strategies for investigating the effectiveness of special education when using large-scale assessment data. Hanushek et al. (1998, 2002) used test scores of students
who transitioned in and out of special education in order to identify an appropriate comparison group. They found that participation in special education programs resulted in an overall increase in student achievement. Ysseldyke and Bielinski (2002) argued for the use of a cohort-static method of identifying the special education group in achievement gap research. When this method was used, the authors found that participation in special education resulted in a decrease in the achievement gap between students with and without disabilities. The current study will seek to replicate and expand upon these findings with a more representative sample of special education students. The current study will provide additional data on special education effectiveness by using these data-analytic techniques when examining the scores of North Carolina students with disabilities on a statewide test of achievement.
CHAPTER THREE

Research Aims

Statement of the Problem

Special education is an extra resource for students with disabilities that is designed to improve academic outcomes for these students. Despite the stated goals of special education, some research suggests that the achievement gap between special education students and general education students continues to grow as students progress through school. Furthermore, current research indicates that students with disabilities still struggle academically several decades after the passage of PL 94-142. These factors combined with popular press demands to know the benefit of continued expenditures for special education have resulted in a need for researchers to find appropriate ways of measuring the effectiveness of special education.

Historically, the effectiveness of special education has been addressed by examining outcomes for students in special education programs. Determinations of what students should be considered “special education students” were often made on a year-to-year basis. That is, group affiliation was determined according to each child’s special education status for the year the outcome data were collected. These examinations typically compared academic outcomes of students with disabilities to students without disabilities, student outcomes across placement types (e.g., outcomes for students who receive their education in separate settings to outcomes of students who receive their education in general education settings), and student outcomes across disability type. A review of the literature indicates that because special education effectiveness is such a large overall question, and because investigations of special education effectiveness are fraught with methodological limitations,
it is important to take multiple methodological approaches in order to address pieces of the special education effectiveness puzzle. Two methodological limitations that are common in special education effectiveness research, (a) identifying special education group affiliation and (b) the use of inappropriate comparison groups, were addressed in the current study.

Defining group composition is an important but often overlooked aspect of special education effectiveness research. In any given year, students who exit special education are a higher achieving group than those who enter special education. Therefore, over time, students who are receiving special education services become a more concentrated group of low-achieving students (Ysseldyke & Bielinski, 2002). When evaluating special education effectiveness, it is important to consider students who may have been helped by special education services in the past, even if they are not currently receiving services. One way of including this group of students would be to define the special education group membership in one academic year, and then follow the progress of that group of students as they move through subsequent academic years without changing their group status. In this method, called the cohort-static method, the highest achieving special education students (those most likely to exit special education services) would continue to represent special education even after they no longer receive services.

One way of addressing the problem of inadequate comparison groups would be to examine students’ performance over multiple data points on large-scale assessments. In this type of within-subjects design, researchers can compare the effectiveness of special education by comparing student performance in years they received special education services to years they did not receive services. Because most students now take part in large-
scale assessments, tracking the progress of students with disabilities in this way is in line with current trends in education policy.

The current study used North Carolina End of Grade Assessments to investigate special education effectiveness. First, the study attempted to replicate the special education effectiveness study of Ysseldyke and Bielinski (2002), who found that the achievement gap between special and regular education students does not continue to grow when a cohort-static method is used to compare the groups. Second, as proposed by Ysseldyke and Bielinski (2002), the performance on large-scale assessments of the select group of students who transition in and out of special education was analyzed using an ABA design. Students who transition in and out of special education tend to be a high performing group (relative to special education students in general), and from a policy perspective this group is most affected by decisions to expand or contract special education services (Hanushek et al., 2002). Therefore, it is important to investigate the effectiveness of special education for this subgroup of special education students.

The current study used data collected from elementary and middle school students in North Carolina. Most North Carolina students in grades three through eight participate in End of Grade Reading (EOG-R) and Mathematics (EOG-M) tests (PSNC, 1998). The North Carolina Department of Public Instruction (NCDPI) has maintained data files of the EOG scores for the state’s students. The files are separated by grade and are currently housed at Duke University. The reading and math test scores within these files served as the dependent variables in the current study.

The results of the current study have the potential to influence the field of special education in two important ways. First, the current study will either support or challenge the
Special Education Effectiveness

findings of Ysseldyke and Bielinski (2002). Given that many students with disabilities were excluded from large-scale assessments in the Ysseldyke and Bielinski study, the current study will seek to improve on the results by using a sample that is more representative of the special education population in North Carolina. Second, the use of a within subjects design allowed for an investigation of special education effectiveness that was not limited by inadequate comparison groups. Taken together, the proposed study has the potential to address the question of the impact of special education resources on the academic achievement of students with disabilities.

Research Hypotheses

Hypothesis 1: The achievement gap between regular and special education as measured by the EOG-R test in sixth grade will be larger when calculated using the cohort-dynamic method than when the cohort-static method is used.

Hypothesis 2: The achievement gap between regular and special education as measured by the EOG-R test in eighth grade will be larger when calculated using the cohort-dynamic method than when the cohort-static method is used.

Hypothesis 3: The achievement gap between regular and special education as measured by the EOG-M test in sixth grade will be larger when calculated using the cohort-dynamic method than when the cohort-static method is used.

Hypothesis 4: The achievement gap between regular and special education as measured by the EOG-M test in eighth grade will be larger when calculated using the cohort-dynamic method than when the cohort-static method is used.

These hypotheses were tested by plotting the achievement gap as an effect size using the cohort-dynamic and cohort-static methods. For each year, EOG test scores of general
education students were used as the reference group, and EOG test scores for special education students were the focal group. In the cohort-static method, the reference group were students who were not receiving special education services in fourth grade. In the cohort-dynamic method, group affiliation was determined according to students’ special education status in each analyzed year. That is, the groups were reclassified in 4th, 6th and 8th grades. Therefore, the reference group in the cohort-dynamic method was all students who did not receive special education services in the specified year, and the focal group was all students who did receive services in the specified year.

**Hypothesis 5:** Students’ academic gain as measured by the EOG-R tests will be greater the year they were in special education than the year before and the year after they were in special education.

**Hypothesis 6:** Students’ academic gain as measured by the EOG-M tests will be greater the year they were in special education than the year before and the year after they were in special education.

As first proposed by Ysseldyke and Bielinski (2002), it was hypothesized that the EOG residual gain scores for students would be greatest in the year they received special education services. Only the achievement of students who transition in and out of special education in consecutive years will be analyzed. The first data-point was the year before the student entered special education, the second data-point was the year they received services, and the third-data point was the year they exited special education.
CHAPTER FOUR

Method

The current study used large-scale assessment data to investigate the effectiveness of special education programs. Extant datasets housed at the North Carolina Education Research Data Center were used. The data files, provided by the North Carolina Department of Public Instruction contain annual test scores for students from public and private schools in North Carolina. The files were assessed after receiving permission from the North Carolina State University Institutional Review Board.

Participants

The populations from which participants were drawn for the Hypotheses 1-4 versus 5-6 differed considerably. Therefore, participants for the first four hypotheses are described in detail, followed by the participants for the remaining hypotheses.

Participants for Hypotheses 1 and 2 included students who (a) took the North Carolina EOG-R test as fourth grade students in either 1998 or 1999, (b) took the EOG-R test again two years later as sixth grade students in 2000 or 2001, (c) took the EOG-R test again two years later as eighth grade students in 2002 or 2003, and (d) had scores reported within the North Carolina Education Research Data Center (NCERDC) database.

Participants for Hypotheses 3 and 4 included students who (a) took the North Carolina EOG-M test as fourth grade students in either 1998 or 1999, (b) took the EOG-M test again two years later as sixth grade students in 2000 or 2001, (c) took the EOG-M test again two years later as eighth grade students in 2002 or 2003, and (d) had scores reported within the North Carolina Education Research Data Center (NCERDC) database. Table 1 presents demographic data for the cohorts that were used to investigate hypotheses 1-4.
The North Carolina Department of Public Instruction identifies twelve classifications of student disability: (a) not identified as a student with a disability, (b) academically gifted, (c) behaviorally-emotionally handicapped, (d) hearing impaired, (e) educable mentally handicapped, (f) specific learning disability, (g) speech-language impaired, (h) visual impairment, (i) other health impaired, (j) orthopedic impairment, (k) traumatic brain injury, and (l) other exceptional classification (PSNC, 1999b). For the current study, students in the first two categories were combined to represent the general education students. The students in the other ten categories were combined to represent the special education students (see Table 1 for percentage of participants by disability category).

Participants for Hypothesis 5 and 6 included North Carolina elementary students (fourth through seventh grade) who participated in the EOG-R (Hypothesis 5) or EOG-M (Hypothesis 6) tests for four consecutive years between 1999 and 2003 and had scores for all four years reported within the NCERDC database (although student scores were only tracked for three consecutive year, students needed to have four years worth of data so that residual gain scores could be calculated for the first year of interest). Four groups of students who were in general education one year, moved into special education the next, and moved back out of special education the next year were selected for each hypothesis. Two groups comprised students who were not in special education in fourth grade, then transitioned into special education in fifth grade and transitioned back out of special education in sixth grade (one group was students who were in fourth grade in 2000, the second was students who were in fourth grade in 2001). Similarly, the other two groups were comprised of students who were not in special education in fifth grade in
either 2000 or 2001, then transitioned into and back out of special education in sixth and seventh grades (Figure 2). To investigate Hypotheses 5 and 6, the cohorts were then combined such that

Table 1

Beginning Demographic Characteristics of the Sample for Hypotheses 1-4

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>36,890 52.0</td>
<td>38,527 52.1</td>
<td>36,952 51.9</td>
<td>38,629 51.9</td>
</tr>
<tr>
<td>Male</td>
<td>34,059 48.0</td>
<td>35,489 47.9</td>
<td>34,317 48.1</td>
<td>35,763 48.1</td>
</tr>
<tr>
<td>Missing</td>
<td>3 .0</td>
<td>0 .0</td>
<td>3 .0</td>
<td>0 .0</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>46,880 66.1</td>
<td>48,143 65.0</td>
<td>47,029 66.0</td>
<td>48,303 64.9</td>
</tr>
<tr>
<td>Black</td>
<td>19,856 28.0</td>
<td>20,946 28.3</td>
<td>19,995 28.1</td>
<td>21,115 28.4</td>
</tr>
<tr>
<td>Hispanic</td>
<td>1,398 2.0</td>
<td>1,755 2.4</td>
<td>1,415 2.0</td>
<td>1,783 2.4</td>
</tr>
<tr>
<td>American Indian</td>
<td>1,048 1.5</td>
<td>1,058 1.4</td>
<td>1,060 1.5</td>
<td>1,062 1.4</td>
</tr>
<tr>
<td>Asian</td>
<td>1,032 1.4</td>
<td>1,175 1.6</td>
<td>1,033 1.4</td>
<td>1,185 1.6</td>
</tr>
<tr>
<td>Other</td>
<td>738 1.0</td>
<td>939 1.3</td>
<td>734 1.0</td>
<td>944 1.3</td>
</tr>
<tr>
<td>Missing</td>
<td>6 .0</td>
<td>0 .0</td>
<td>6 .0</td>
<td>0 .0</td>
</tr>
<tr>
<td><strong>Special education status</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>64,189 90.5</td>
<td>66,890 90.4</td>
<td>64,222 90.1</td>
<td>66,936 90.0</td>
</tr>
<tr>
<td>Special</td>
<td>6,763 9.5</td>
<td>7,126 9.6</td>
<td>7,050 9.9</td>
<td>7,456 10.0</td>
</tr>
<tr>
<td><strong>Disability</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLD</td>
<td>4,236 6.0</td>
<td>4,421 6.0</td>
<td>4,493 6.3</td>
<td>4,691 6.3</td>
</tr>
<tr>
<td>SLI</td>
<td>1,009 1.4</td>
<td>1,036 1.4</td>
<td>1,015 1.4</td>
<td>1,049 1.4</td>
</tr>
<tr>
<td>BEH</td>
<td>404 .6</td>
<td>418 .6</td>
<td>410 .6</td>
<td>430 .6</td>
</tr>
<tr>
<td>OHI</td>
<td>584 .8</td>
<td>664 .9</td>
<td>592 .8</td>
<td>671 .9</td>
</tr>
<tr>
<td>Other</td>
<td>530 .7</td>
<td>587 .8</td>
<td>540 .8</td>
<td>615 .8</td>
</tr>
<tr>
<td><strong>Free or Reduced Lunch</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>70,052 42.5</td>
<td>74,016 42.0</td>
<td>72,172 42.5</td>
<td>74,392 41.9</td>
</tr>
</tbody>
</table>
there was one reading sample and one math sample, thus allowing for analyses of EOG-R and EOG-M scores for the year students were in special education to the years they were not in special education across grade levels and years. Because all of the students who met selection criteria for these hypotheses took both the EOG-R and EOG-M tests in the selected years, the compositions of the two groups are identical. Demographic information for both individual years and the total sample used to investigate Hypotheses 5 and 6 is presented in Table 2.

<table>
<thead>
<tr>
<th>Student Groups</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2000</td>
</tr>
<tr>
<td>Students who entered 4th grade in 2000</td>
<td>4th grade general education</td>
</tr>
<tr>
<td>Students who entered 4th grade in 2001</td>
<td>5th grade general education</td>
</tr>
<tr>
<td>Students who entered 5th grade in 2000</td>
<td>5th grade general education</td>
</tr>
<tr>
<td>Students who entered 5th grade in 2001</td>
<td>5th grade general education</td>
</tr>
</tbody>
</table>

Figure 2. Student groups by year and grade level.
Table 2

Demographic Characteristics of the Sample for Hypotheses 5-6 (Students With Multiple Transitions Between Special Education and General Education)

<table>
<thead>
<tr>
<th></th>
<th>2000 4th Grade</th>
<th>2001 4th Grade</th>
<th>2000 5th Grade</th>
<th>2001 5th Grade</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
<td>N</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>128</td>
<td>38.4</td>
<td>139</td>
<td>40.8</td>
<td>87</td>
</tr>
<tr>
<td>Male</td>
<td>205</td>
<td>61.6</td>
<td>202</td>
<td>59.2</td>
<td>138</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>218</td>
<td>65.5</td>
<td>201</td>
<td>58.9</td>
<td>142</td>
</tr>
<tr>
<td>Black</td>
<td>94</td>
<td>28.2</td>
<td>107</td>
<td>31.4</td>
<td>69</td>
</tr>
<tr>
<td>Hispanic</td>
<td>7</td>
<td>2.1</td>
<td>18</td>
<td>5.3</td>
<td>1</td>
</tr>
<tr>
<td>American Indian</td>
<td>2</td>
<td>0.6</td>
<td>6</td>
<td>1.8</td>
<td>10</td>
</tr>
<tr>
<td>Asian</td>
<td>7</td>
<td>2.1</td>
<td>4</td>
<td>1.2</td>
<td>2</td>
</tr>
<tr>
<td>Other</td>
<td>5</td>
<td>1.5</td>
<td>5</td>
<td>1.5</td>
<td>1</td>
</tr>
<tr>
<td>Disability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLD</td>
<td>118</td>
<td>35.4</td>
<td>76</td>
<td>22.3</td>
<td>99</td>
</tr>
<tr>
<td>SLI</td>
<td>162</td>
<td>48.7</td>
<td>189</td>
<td>55.4</td>
<td>57</td>
</tr>
<tr>
<td>BEH</td>
<td>8</td>
<td>2.4</td>
<td>6</td>
<td>1.8</td>
<td>20</td>
</tr>
<tr>
<td>OHI</td>
<td>23</td>
<td>6.9</td>
<td>42</td>
<td>12.3</td>
<td>21</td>
</tr>
<tr>
<td>Hearing</td>
<td>14</td>
<td>4.2</td>
<td>10</td>
<td>2.9</td>
<td>15</td>
</tr>
<tr>
<td>Other</td>
<td>8</td>
<td>2.4</td>
<td>18</td>
<td>5.3</td>
<td>13</td>
</tr>
<tr>
<td>Total</td>
<td>333</td>
<td>29.1</td>
<td>341</td>
<td>29.9</td>
<td>225</td>
</tr>
</tbody>
</table>

Measures

The outcome variables for all analyses were calculated from students’ scores on the EOG-R and EOG-M tests. The tests are designed to measure students’ progress and how well they were able to meet academic achievement requirements in reading and mathematics. Options for score reporting are percentiles, developmental scale scores, and achievement
levels. The developmental scale scores are calculated on a continuous, equal interval scale that allows for comparison of scores across grade levels, although separate developmental scales are used for reading and mathematics.

Given that the analyses for Hypotheses 1-4 were used to examine the achievement gap across multiple, non-consecutive academic years, developmental scale scores that had been converted to an effect size to represent the achievement gap were used as the outcome variable (Boyd et al., 2008). Therefore, the outcome variable for Hypotheses 5 and 6 analyses were the students’ residual gain scores on the EOG-R and EOG-M tests. Residual gain scores take into account students’ scores on the previous year’s test and are useful for tracking student progress in individual academic years (PSNC, 2004).

**EOG-R.** The EOG-R tests assess the ability of students to use strategies that enhance reading comprehension, including reading to acquire, interpret, and apply information, and reading for critical analysis and evaluation. Reading is assessed by having students read passages and then answer questions related to the passages. Ten passages are included on each EOG-R test, including four literary passages, four content-based passages, and two consumer/human interest passages. The number of questions ranges from 56 in fourth grade to 68 in eighth (PSNC, 1999a). The tests measure four broad concepts within the area of reading comprehension: cognition, interpretation, critical stance and connections (PSNC 2004).

Internal consistency reliability estimates for the EOG-R tests (coefficient alphas) range from .92 (8th grade) to .94 (6th grade). The standard error of measurement for the EOG-
R tests is two to six points. Thus, the EOG-R tests demonstrate adequate reliability (PSNC, 2004).

For the North Carolina EOG-R tests, evidence of validity is provided through content relevance, response processes, relationship of test scores to other external variables, and maintaining consistency in the testing environment. Therefore, evidence of both content and criterion validity exists for the EOG-R tests. Content relevance is demonstrated through analyses of the judged match between item content and the constructs (or strands) emphasized in the curriculum: cognition, interpretation, critical stance, and connections. The percentages of questions that measure each strand are reported at every grade level. For the fourth through eighth grade EOG-R tests, 26-39% of the questions measure cognition, 37-42% of the questions measure interpretation, 18-27% measure critical stance and 4-6% measure connections (PSNC, 2004).

Teachers’ judgments of the relationship between the EOG-R tests and their associated variables (student achievement, expected grade, and assigned achievement levels) serve as another source of evidence of the tests’ validity. Pearson correlation coefficients between the EOG-R tests and teacher judgments of associated variables range from 0.46 to 0.65, indicating moderate to strong correlations between scale scores and their associated variables (PSNC, 2004).

EOG-M. The EOG-M test assesses students’ abilities in the seven strands of the mathematics curriculum: numeration, geometry, patterns and pre-algebra, measurement, problem-solving, data analysis and statistics, and computation. The mathematics test is administered in two parts: mathematics computation (calculator use not allowed) and mathematics applications (calculator use allowed). The mathematics computation section
Special Education Effectiveness

consists of twelve questions in fourth through sixth grades, and eight questions in seventh and eighth grades. The mathematics applications section consists of 68 questions in fourth to sixth grades and 72 questions in seventh and eighth grades (PSNC, 2004).

Internal consistency reliability estimates for the EOG-M tests (coefficient alphas) range from .94 (8th grade) to .96 (6th grade). The standard error of measurement for the EOG-M tests is two to six points (PSNC, 2006). Thus, the EOG-M tests demonstrate adequate reliability.

For the North Carolina EOG-M, evidence of validity is provided through content relevance and the relationship of test scores to other external variables. Content relevance is demonstrated through analyses of the judged match between item content and the constructs (or strands) emphasized in the curriculum: number sense, numeration, and numerical operations; spatial sense, measurement, and geometry; patterns, relationships, and functions; and data, probability, and statistics. Unlike the method used for reporting content relevance for the EOG-R tests (i.e. reporting the number of items that assess each content area), evidence of the EOG-M tests’ content validity is demonstrated through teacher ratings. EOG-M test items are rated by teachers as adhering to the following criteria: (a) the test content reflects the goals and objectives of the grade level curriculum; (b) the test content reflects the goals and objectives of the curriculum as the curriculum is taught in their school or school system; (c) the items are clearly and concisely written, and the vocabulary is appropriate to the target age level; (d) the content is balanced (not biased) in relation to ethnicity, race, sex, socioeconomic status, and geographic districts of the state; and (e) each of the items has one and only one answer that is best; however, the distracters appear plausible for someone who has not achieved mastery of the represented objective. Teacher responses to these
questionnaires indicate that the tests met these criteria to a “high” or “superior” degree (PSNC, 2006).

Teachers’ judgments of the relationship between the EOG-M tests and their associated variables (student achievement, expected grade, and assigned achievement levels) serve as sources of evidence of criterion validity. Pearson correlation coefficients for the EOG-M tests range from 0.49 to 0.70, indicating moderate to strong correlations between scale scores and teacher judgments of students’ achievement (PSNC, 2006).

EOG Test Administration

The North Carolina EOG tests are administered to students in third through eighth grades as part of the statewide assessment program in May of each year. Only school personnel are allowed to administer the tests, and school test coordinators are responsible for monitoring test administrations within the building and responding to situations that may arise during test administrations. All students in third through eighth grades are required to participate in the EOG tests. On a case-by-case basis where appropriate documentation exists, students with disabilities and students with limited English proficiency may receive testing accommodations. The need for accommodations must be documented in a current Individualized Education Program (IEP), Section 504 Plan or LEP Plan. The accommodations must be used routinely during the student’s instructional program or similar classroom assessments (PSNC 1999a). A small percentage of students are exempted from the testing or take an alternate form of the test. However, if students are following the general education curriculum, the expectation is that they will participate in the annual large scale testing in that subject area.
CHAPTER FIVE

Results

This chapter will describe the data analysis procedures and related results for each of the research questions and hypotheses presented in Chapter Three. The chapter begins with a description of the preliminary data preparation steps. The analyses and results related to the hypotheses outlined in Chapter Three are then presented.

Descriptive and Preliminary Analyses

As noted in Chapter Four, the present study made use of extant achievement test data for students in the state of North Carolina to investigate the achievement gap between special education and general education students and to compare student achievement for the years they were in special education to the years they were not in special education. To create reference scores for Hypotheses 1-4, and to create residual gain scores for Hypotheses 5-6, data from all students in the analyzed years needed to be included for preliminary analyses. A description of procedures used to merge the data sets, create effect sizes for Hypotheses 1-4, calculate residual gain scores for Hypotheses 5-6, and descriptive statistics for test scores across years is presented in this section before the analyses specific to each of the hypotheses are presented.

Merging Data Sets and Creating Cohorts

The NCERDC houses student level data for each grade level and year. Given that the intent of the current study was to compare student achievement across years, individual datasets needed to be merged to create the identified cohort groups. For Hypotheses 1-4, this process required combining data sets for the years 1998, 2000, and 2002 as well as combining data sets for 1999, 2001, and 2003. Once these data sets were merged, students
who did not have EOG scores for all three years of interest were eliminated from the data set. The cohort static groups were defined by identifying students’ special education status in the first year of interest (1998 or 1999). Once the cohort static groups were defined, the achievement gap between special and general education students could be calculated for the other years, as group affiliation did not change (by definition) for this group. For the cohort dynamic groups, an extra step needed to be taken at each subsequent data point, as student affiliation may have changed. That is, for the cohort dynamic cohorts, group affiliation was redefined every year according to students’ special education status in that particular year.

A similar procedure was used to merge the data sets in preparation for investigation of Hypotheses 5 and 6. For example, to create one set of cohort data for Hypothesis 5, data for students in 3rd grade in 1999 needed to be merged with data for students in 4th grade in 2000, 5th grade in 2001 and sixth grade in 2002. Once merged, all students who did not have scores for all four years were eliminated from the data set. Next, only students who were in general education in 2000, special education in 2001, and general education in 2002 were retained. This procedure was repeated for students who entered 4th grade in 2001, students who entered 5th grade in 2000, and students who entered 5th grade in 2001. Once the groups were defined, residual gain scores as measured by the EOG-R and EOG-M for the students in each cohort were calculated using the EOG scores of the entire student population across those four years. These residual gain scores were then used for the main analyses in Hypotheses 5 and 6, presented later.

Creating Effect Sizes

To complete the analyses for Hypotheses 1-4, effect sizes for the difference between the students in special and general education needed to be computed for each cohort. Effect
sizes are computed by subtracting the mean score of a focal group (e.g., special education students) from the mean score of a reference group (e.g., general education students) and dividing the difference by the pooled standard deviation. To calculate the effect sizes for each cohort, the means and standard deviations of the EOG-R and EOG-M scores were calculated in the original NCERDC student files by year and grade. The mean EOG scores of the focal (special education) and reference (general education) groups, along with the pooled standard deviations used to calculate the effect sizes are presented in Table 3.

Table 3

<table>
<thead>
<tr>
<th>Means and Pooled Standard Deviations to Compute Effect Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cohorts</strong></td>
</tr>
<tr>
<td><strong>Cohorts</strong></td>
</tr>
<tr>
<td>4th 1998</td>
</tr>
<tr>
<td>6th 2000</td>
</tr>
<tr>
<td>8th 2002</td>
</tr>
<tr>
<td>4th 1999</td>
</tr>
<tr>
<td>6th 2001</td>
</tr>
<tr>
<td>8th 2003</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pooled SD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>4th 1998</strong></td>
</tr>
<tr>
<td><strong>6th 2000</strong></td>
</tr>
<tr>
<td><strong>8th 2002</strong></td>
</tr>
<tr>
<td><strong>4th 1999</strong></td>
</tr>
<tr>
<td><strong>6th 2001</strong></td>
</tr>
<tr>
<td><strong>8th 2003</strong></td>
</tr>
</tbody>
</table>
**Normalizing Raw Test Scores and Creating Residualized Gain Scores**

To complete the analyses for Hypotheses 5 and 6, normalized z-scores needed to be created from the students’ raw test data. Then, to control for regression to the mean, the z-scores were used to create residual gain scores, and these residual gain scores served as the outcome variable for Hypotheses 5 and 6.

To create the z-scores, the means and standard deviations of all students for the EOG-R and EOG-M were calculated in the original NCERDC student files separately by year and grade. The z-scores were then created by subtracting the mean EOG Test score for that year and grade from the students’ individual scores, and dividing the difference by the pooled standard deviation of the appropriate grade and year.

To calculate the residual gain scores, students’ test scores for the year in question (year two) were correlated with their test scores from the previous year (year one). The predicted score for each student was then calculated by multiplying the z-score from year one by the correlation coefficient. The residual gain score was then calculated by subtracting the predicted score from the year two z-score. That is, the differences between the students’ predicted and actual test scores were their residual gain scores. The correlation coefficients used to create the residual gain scores for each cohort and year in question are presented in Table 4. Tables 5 and 6 report the mean z-scores and the residual gain scores on the EOG-R and EOG-M tests for each cohort.
Table 4

Correlations Used to Compute Residual Gain Scores

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Correlation coefficients</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>EOG-R</td>
<td>EOG-M</td>
</tr>
<tr>
<td>Fourth Grade 2000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999-2000</td>
<td>.82</td>
<td>.82</td>
<td></td>
</tr>
<tr>
<td>2000-2001</td>
<td>.82</td>
<td>.84</td>
<td></td>
</tr>
<tr>
<td>2001-2002</td>
<td>.82</td>
<td>.87</td>
<td></td>
</tr>
<tr>
<td>Fourth Grade 2001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000-2001</td>
<td>.81</td>
<td>.82</td>
<td></td>
</tr>
<tr>
<td>2001-2002</td>
<td>.81</td>
<td>.85</td>
<td></td>
</tr>
<tr>
<td>2002-2003</td>
<td>.80</td>
<td>.86</td>
<td></td>
</tr>
<tr>
<td>Fifth Grade 2000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999-2000</td>
<td>.83</td>
<td>.85</td>
<td></td>
</tr>
<tr>
<td>2000-2001</td>
<td>.84</td>
<td>.86</td>
<td></td>
</tr>
<tr>
<td>2001-2002</td>
<td>.85</td>
<td>.88</td>
<td></td>
</tr>
<tr>
<td>Fifth Grade 2001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000-2001</td>
<td>.82</td>
<td>.84</td>
<td></td>
</tr>
<tr>
<td>2001-2002</td>
<td>.82</td>
<td>.87</td>
<td></td>
</tr>
<tr>
<td>2002-2003</td>
<td>.82</td>
<td>.88</td>
<td></td>
</tr>
</tbody>
</table>
Table 5

z-Score Means and Standard Deviations for Students Who Transitioned In and Out of Special Education by Cohort and Year

<table>
<thead>
<tr>
<th>Cohort</th>
<th>EOG-R</th>
<th>EOG-M</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fourth Grade 2000</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999 (third grade)</td>
<td>-0.20 (0.95)</td>
<td>-0.16 (0.94)</td>
</tr>
<tr>
<td>2000 (fourth grade)</td>
<td>-0.20 (0.94)</td>
<td>-0.21 (0.92)</td>
</tr>
<tr>
<td>2001 (fifth grade)</td>
<td>-0.19 (0.98)</td>
<td>-0.16 (0.96)</td>
</tr>
<tr>
<td>2002 (sixth grade)</td>
<td>-0.18 (0.99)</td>
<td>-0.12 (0.96)</td>
</tr>
<tr>
<td><strong>Fourth Grade 2001</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000 (third grade)</td>
<td>-0.20 (0.96)</td>
<td>-0.10 (0.97)</td>
</tr>
<tr>
<td>2001 (fourth grade)</td>
<td>-0.19 (0.92)</td>
<td>-0.17 (0.95)</td>
</tr>
<tr>
<td>2002 (fifth grade)</td>
<td>-0.16 (0.99)</td>
<td>-0.13 (0.96)</td>
</tr>
<tr>
<td>2003 (sixth grade)</td>
<td>-0.18 (1.02)</td>
<td>-0.09 (0.94)</td>
</tr>
<tr>
<td><strong>Fifth Grade 2000</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999 (fourth grade)</td>
<td>-0.21 (0.94)</td>
<td>-0.16 (0.95)</td>
</tr>
<tr>
<td>2000 (fifth grade)</td>
<td>-0.22 (0.96)</td>
<td>-0.16 (0.97)</td>
</tr>
<tr>
<td>2001 (sixth grade)</td>
<td>-0.16 (0.93)</td>
<td>-0.13 (0.91)</td>
</tr>
<tr>
<td>2002 (seventh grade)</td>
<td>-0.20 (0.98)</td>
<td>-0.13 (0.95)</td>
</tr>
<tr>
<td><strong>Fifth Grade 2001</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000 (fourth grade)</td>
<td>-0.14 (1.03)</td>
<td>-0.09 (1.01)</td>
</tr>
<tr>
<td>2001 (fifth grade)</td>
<td>-0.16 (1.05)</td>
<td>-0.08 (1.00)</td>
</tr>
<tr>
<td>2002 (sixth grade)</td>
<td>-0.17 (1.06)</td>
<td>-0.10 (1.02)</td>
</tr>
<tr>
<td>2004 (seventh grade)</td>
<td>-0.05 (1.04)</td>
<td>-0.02 (1.04)</td>
</tr>
</tbody>
</table>

52
Table 6

Residual Gain Scores and Standard Deviations for Students Who Transitioned In and Out of Special Education by Cohort and Year

<table>
<thead>
<tr>
<th>Cohort</th>
<th>EOG-R</th>
<th>EOG-M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fourth Grade 2000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000 (fourth grade, A1)</td>
<td>-0.04 (0.57)</td>
<td>-0.08 (0.58)</td>
</tr>
<tr>
<td>2001 (fifth grade, B)</td>
<td>-0.04 (0.58)</td>
<td>0.02 (0.58)</td>
</tr>
<tr>
<td>2002 (sixth grade, A2)</td>
<td>-0.02 (0.59)</td>
<td>0.02 (0.51)</td>
</tr>
<tr>
<td>Fourth Grade 2001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001 (fourth grade, A1)</td>
<td>-0.03 (0.54)</td>
<td>-0.08 (0.56)</td>
</tr>
<tr>
<td>2002 (fifth grade, B)</td>
<td>-0.01 (0.59)</td>
<td>0.01 (0.55)</td>
</tr>
<tr>
<td>2003 (sixth grade, A2)</td>
<td>-0.02 (0.56)</td>
<td>0.02 (0.49)</td>
</tr>
<tr>
<td>Fifth Grade 2000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000 (fifth grade, A1)</td>
<td>-0.04 (0.56)</td>
<td>-0.02 (0.53)</td>
</tr>
<tr>
<td>2001 (sixth grade, B)</td>
<td>-0.03 (0.59)</td>
<td>0.00 (0.54)</td>
</tr>
<tr>
<td>2002 (seventh grade, A2)</td>
<td>-0.06 (0.59)</td>
<td>-0.01 (0.49)</td>
</tr>
<tr>
<td>Fifth Grade 2001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001 (fifth grade, A1)</td>
<td>-0.04 (0.57)</td>
<td>0.00 (0.53)</td>
</tr>
<tr>
<td>2002 (sixth grade, B)</td>
<td>-0.04 (0.57)</td>
<td>-0.03 (0.48)</td>
</tr>
<tr>
<td>2003 (seventh grade, A2)</td>
<td>0.09 (0.59)</td>
<td>0.07 (0.52)</td>
</tr>
<tr>
<td>Cohorts Combined</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>-0.03 (0.60)</td>
<td>-0.05 (0.55)</td>
</tr>
<tr>
<td>B</td>
<td>-0.02 (0.59)</td>
<td>0.00 (0.54)</td>
</tr>
<tr>
<td>A2</td>
<td>-0.01 (0.58)</td>
<td>0.02 (0.50)</td>
</tr>
</tbody>
</table>

Analyses Specific to Hypotheses

Comparing Methods of Reporting

Hypotheses 1 through 4, as described in Chapter Four, asserted that the achievement gap between general and special education would be larger at each selected time-point when calculated using the cohort-dynamic method than when the cohort-static method is used. Specifically, it was expected that the achievement gap, measured as an effect size for special education, would continue to grow when measured using the cohort dynamic method.
Additionally, the difference between the two methods of identifying student status was expected to yield significantly different effect sizes at each selected time-point, as evidenced by non-overlapping confidence intervals for the effect sizes. The confidence intervals were calculated using the sample means and the pooled standard deviation of the samples. The standard error was computed by dividing the pooled standard deviation by the sample size. The confidence interval was then determined for each sample’s test scores by multiplying 1.96 (the z-score value) and the standard error, and then subtracting this number from the mean for that sample. Tables 7 and 8 provide the effect sizes and confidence intervals for special education for the two cohorts examined (1998 and 1999) as measured by the EOG-R and EOG-M.

Overall, non-overlapping confidence intervals were observed at every time interval between the two methods of reporting group affiliation for both cohorts. For example, for the cohort that consisted of students who were in fourth grade in 1998, using the Cohort Static method of reporting student affiliation, the effect size representing the achievement gap on the EOG-R in sixth grade was -.91, with a 95% confidence interval from -.93 to -.89. Using the same cohort and the Cohort Dynamic method of reporting student affiliation, the effect size representing the achievement gap on the EOG-R in sixth grade was -1.06, with a 95% confidence interval of -1.08 to -1.04. Therefore, it can be said with 95% certainty that the two methods of reporting student affiliation for this cohort resulted in different achievement performance gaps (stated as an effect sizes).
Table 7

The Effect Sizes and Confidence Intervals Across Grades Between General Education and Special Education Students for Students who were in Fourth Grade in 1998

<table>
<thead>
<tr>
<th>Dynamic</th>
<th>EOG-R</th>
<th>EOG-R</th>
<th>EOG-M</th>
<th>EOG-M</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cohort Static</td>
<td>Cohort Dynamic</td>
<td>Cohort Static</td>
<td>Cohort</td>
</tr>
<tr>
<td>4th Grade 1998</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ES (CI)</td>
<td>$d = -1.01$</td>
<td>$d = -1.01$</td>
<td>$d = -.84$</td>
<td>$d = -.84$</td>
</tr>
<tr>
<td>6th Grade 2000</td>
<td>($-1.03$ to $-.99$)</td>
<td>($-1.03$ to $-.99$)</td>
<td>($-.86$ to $-.82$)</td>
<td>($-.86$ to $-.82$)</td>
</tr>
<tr>
<td>ES (CI)</td>
<td>$d = -.91$</td>
<td>$d = -1.07$</td>
<td>$d = -.86$</td>
<td>$d = -1.00$</td>
</tr>
<tr>
<td>8th Grade 2002</td>
<td>($-.93$ to $-.89$)</td>
<td>($-1.09$ to $-1.03$)</td>
<td>($-.88$ to $-.84$)</td>
<td>($-1.02$ to $-.98$)</td>
</tr>
<tr>
<td>ES (CI)</td>
<td>$d = -.85$</td>
<td>$d = -1.02$</td>
<td>$d = -.82$</td>
<td>$d = -1.00$</td>
</tr>
<tr>
<td></td>
<td>($-.87$ to $-.83$)</td>
<td>($-1.04$ to $-1.00$)</td>
<td>($-.84$ to $-.80$)</td>
<td>($-1.02$ to $-.98$)</td>
</tr>
</tbody>
</table>

Table 8

The Effect Sizes and Confidence Intervals Across Grades Between General Education and Special Education Students for Students who were in Fourth Grade in 1999

<table>
<thead>
<tr>
<th>Dynamic</th>
<th>EOG-R</th>
<th>EOG-R</th>
<th>EOG-M</th>
<th>EOG-M</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cohort Static</td>
<td>Cohort Dynamic</td>
<td>Cohort Static</td>
<td>Cohort</td>
</tr>
<tr>
<td>4th Grade 1999</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$d = -1.00$ ($-1.02$ to $-.98$)</td>
<td>$d = -1.00$ ($-1.02$ to $-.98$)</td>
<td>$d = -.86$ ($-.88$ to $.84$)</td>
<td>$d = -.86$ ($-.88$ to $.84$)</td>
<td></td>
</tr>
<tr>
<td>6th Grade 2001</td>
<td>$d = -.93$ ($-.95$ to $-.91$)</td>
<td>$d = -1.06$ ($-1.08$ to $-1.04$)</td>
<td>$d = -.90$ ($-.92$ to $.88$)</td>
<td>$d = -1.00$ ($-1.02$ to $-.98$)</td>
</tr>
<tr>
<td>8th Grade 2003</td>
<td>$d = -.85$ ($-.87$ to $-.83$)</td>
<td>$d = -1.03$ ($-1.05$ to $-1.01$)</td>
<td>$d = -.86$ ($-.88$ to $.84$)</td>
<td>$d = -1.03$ ($-1.05$ to $-1.01$)</td>
</tr>
</tbody>
</table>

Similar to the methodology used by Ysseldyke and Bielinski (2002) to show the extent that reclassification affects achievement gap trends, the achievement gap between
special and general education was plotted as an effect size using the Cohort Static and Cohort Dynamic methods. The trend in achievement across five years (4th, 6th and 8th grades) is illustrated in Figures 3 through 6. The x-axis represents the year the test was given, and the y-axis represents the achievement gap between special and general education students defined by the effect size (the horizontal line equal to zero represents the achievement of general education students). Given that the starting point for the current analyses was fourth grade for the cohorts, the student affiliation for the Cohort Static method was determined in fourth grade. Therefore, the Cohort Static and Cohort Dynamic trend lines consisted of identical groups in fourth grade.

Figure 3. Trend line comparison for EOG-R gain scores for students in the 1998 cohort.
Figure 4. Trend line comparison for EOG-R gain scores for students in the 1999 cohort.

Figure 5. Trend line comparison for EOG-M gain scores for students in the 1998 cohort.
In sum, the results of the present study replicated the findings of Ysseldyke and Bielinski study (2002), and provided support for Hypotheses 1-4. The mean achievement for students in special education was well below the mean achievement for students in general education regardless of the reporting method. However, using the Cohort Static method, students in special education either gained ground (EOG-R) or maintained the same achievement gap across time (EOG-M). Using the Cohort Dynamic method, students in special education either maintained the achievement gap (EOG-R) or lost ground to general education students (EOG-M) (see Tables 7 and 8 for the effect sizes for each cohort and method of reporting).

**ABA Analyses**

Hypotheses 5 and 6 asserted that students’ academic gain as measured by the EOG-R and EOG-M tests would be greater the year they were in special education than the year before and the year after they were in special education. Only students who transitioned in
and out of special education in consecutive years and had test scores from four consecutive years were used in these analyses. The students’ residual gain scores, or the amount of academic growth that the students made in one academic year, were used as the basis for these analyses. For these ABA analyses, the residual gain for the year before students were in special education (A) was compared to the residual gain score for the year they were in special education (B). Also, the residual gain score for the year they were in special education (B) was compared to the residual gain score for the year after they exited special education (A). Table 6 provides the mean residual gain scores for each cohort.

To test Hypotheses 5 and 6, a one-way, repeated-measures ANOVA was computed using students’ EOG-R or EOG-M residual gain scores to determine if there was an effect for time. The main analyses for these hypotheses were computed after combining EOG-R scores for all cohorts and EOG-M scores for all cohorts. It was expected that the mean would be greater in the years students received special education services (B) than the year before they received services (A1) and the year after they received services (A2). Therefore, it was expected that these analyses would provide evidence that special education was effective in helping improve students’ academic gain for the year they received services.

When the cohorts were combined, a significant effect for time was found for the EOG-M, \((F(2, 1139) = 5.15, \ p < .01)\) but not for the EOG-R, \((F(2, 1139) = .31, \ p > .05)\). Given the significant main effect for time in the EOG-M analysis, follow-up simple contrasts were performed to compare the mean for the year students received special education (B) to the year before students entered special education (A1) and the year after they exited special education (A2). As predicted in Hypothesis 6, the EOG-M mean score for the year students were in special education was significantly higher than the mean for the year before they
entered special education ($F(2, 1139) = 4.14, p < .05$). However, the mean mathematics score for the year students were in special education and the year after they exited did not differ ($F(2, 1139) = .78, p > .05$). In sum, Hypothesis 5 was not confirmed. The mean reading score for the year students were in special education did not differ from the means for the year before they entered special education or the year they exited again into general education. Hypothesis 6 was partially confirmed. The mean mathematics score for the year students were in special education was higher than the year before they entered special education. However, the mean for the year in special education and following year in general education did not differ.

Although the combined scores of cohorts served as the basis for the hypothesis testing, post hoc analyses were also conducted. Individual cohorts were compared to investigate the nature of any observed differences, identify potential starting group differences that would have threatened the results of the current study, and investigate the potential for individual cohort differences in the absence of combined cohort findings. Results revealed that only three of the eight individual cohort comparisons yielded a significant effect for time (Table 9).

To further investigate the nature of the effect for time and to attempt to identify any change in mean scores at each transition point, students’ EOG scores for the year they received special education services were compared separately to the year before they entered special education (A1) and the year after they exited special education (A2) using $t$-tests (Table 10). Of these 16 comparisons, only three yielded significant results. Two of these indicated that the residual gain scores for the year students were in special education were superior to the year before they entered special education (EOG-M scores for students in the
4th grade 2000 and 4th grade 2001 cohorts). The other significant difference was found between EOG-R scores for students in the 5th grade 2001 cohort, which indicated means were higher for students the year after they exited special education than the year they were in special education. (EOG-R scores for students in sixth and seventh grades in 2001 and 2002).

Table 9

F-values for Repeated Measures ANOVA by Cohort and Test Type

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Sample Size</th>
<th>EOG-R</th>
<th>EOG-M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fourth Grade 2000</td>
<td>333</td>
<td>0.05</td>
<td>2.77</td>
</tr>
<tr>
<td>Fourth Grade 2001</td>
<td>341</td>
<td>0.06</td>
<td>3.25*</td>
</tr>
<tr>
<td>Fifth Grade 2000</td>
<td>225</td>
<td>1.22</td>
<td>0.12</td>
</tr>
<tr>
<td>Fifth Grade 2001</td>
<td>242</td>
<td>3.62*</td>
<td>2.02</td>
</tr>
<tr>
<td>Cohorts Combined</td>
<td>1141</td>
<td>0.31</td>
<td>5.15**</td>
</tr>
</tbody>
</table>

*p < .05, **p < .01

Table 10

Comparison of EOG Gain Scores Between the Year Students Were in Special Education (B) to the Year Before They Entered Special Education (A1) and to the Year After They Exited Special Education (A2)

<table>
<thead>
<tr>
<th>Cohort</th>
<th>t-Score</th>
<th>EOG-R</th>
<th>EOG-M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fourth Grade 2000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1 to B</td>
<td>0.18</td>
<td></td>
<td>2.07*</td>
</tr>
<tr>
<td>B to A2</td>
<td>0.17</td>
<td></td>
<td>0.10</td>
</tr>
<tr>
<td>Fourth Grade 2001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1 to B</td>
<td>0.32</td>
<td></td>
<td>1.90*</td>
</tr>
<tr>
<td>B to A2</td>
<td>0.22</td>
<td></td>
<td>0.26</td>
</tr>
<tr>
<td>Fifth Grade 2000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1 to B</td>
<td>1.22</td>
<td></td>
<td>0.93</td>
</tr>
<tr>
<td>B to A2</td>
<td>1.60</td>
<td></td>
<td>0.77</td>
</tr>
<tr>
<td>Fifth Grade 2001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1 to B</td>
<td>0.04</td>
<td></td>
<td>0.60</td>
</tr>
<tr>
<td>B to A2</td>
<td>2.84*</td>
<td></td>
<td>1.04</td>
</tr>
</tbody>
</table>

*p < .05
To ensure that the findings were not a result of starting differences between cohorts, means for the year before students entered special education were compared between cohorts. Results revealed no significant difference between the ‘A1’ means between the cohorts as measured by the EOG-R and EOG-M (Table 11). Therefore, it is unlikely that findings were the result of baseline differences between cohorts.

Table 11
Comparison of EOG Gain Scores Between Cohorts for the Year Before Students Received Special Education Services (‘A1’)

<table>
<thead>
<tr>
<th>Comparison</th>
<th>EOG-R</th>
<th>EOG-M</th>
</tr>
</thead>
<tbody>
<tr>
<td>4th Grade 2000 to 4th Grade 2001</td>
<td>1.18</td>
<td>0.11</td>
</tr>
<tr>
<td>4th Grade 2000 to 5th Grade 2000</td>
<td>0.53</td>
<td>1.13</td>
</tr>
<tr>
<td>4th Grade 2000 to 5th Grade 2001</td>
<td>0.60</td>
<td>1.66</td>
</tr>
<tr>
<td>4th Grade 2001 to 5th Grade 2000</td>
<td>0.07</td>
<td>1.27</td>
</tr>
<tr>
<td>4th Grade 2001 to 5th Grade 2001</td>
<td>0.01</td>
<td>1.81</td>
</tr>
<tr>
<td>5th Grade 2000 to 5th Grade 2001</td>
<td>0.06</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Cohorts were also compared to determine if there were differences between cohorts in terms of students’ scores for the year they received special education. That is, the ‘B’ means were compared across cohorts to investigate if there were differences for students based on the year and grade they received special education. Results revealed that there were no significant differences between the cohort means on either the EOG-R or EOG-M for the year they received special education services (Table 12). In addition, residual gain scores for students in cohorts that received special education in fifth grade were combined and compared to students who received special education in sixth grade. Results revealed no
significant difference between mean scores based on the grade that they received special education services (Table 12).

<table>
<thead>
<tr>
<th>Comparison</th>
<th>EOG-R</th>
<th>EOG-M</th>
</tr>
</thead>
<tbody>
<tr>
<td>5th Grade 2001 to 5th Grade 2002</td>
<td>0.11</td>
<td>0.14</td>
</tr>
<tr>
<td>5th Grade 2001 to 6th Grade 2001</td>
<td>0.93</td>
<td>0.27</td>
</tr>
<tr>
<td>5th Grade 2001 to 6th Grade 2002</td>
<td>0.41</td>
<td>1.00</td>
</tr>
<tr>
<td>5th Grade 2002 to 6th Grade 2001</td>
<td>1.01</td>
<td>0.16</td>
</tr>
<tr>
<td>5th Grade 2002 to 6th Grade 2002</td>
<td>0.30</td>
<td>0.91</td>
</tr>
<tr>
<td>6th Grade 2001 to 6th Grade 2002</td>
<td>1.25</td>
<td>0.69</td>
</tr>
<tr>
<td>5th Grade Combined to 6th Grade Combined</td>
<td>0.42</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Results of the current study provided partial support for Hypothesis 6, which asserted that for students who transitioned into and back out of special education in consecutive years, EOG-M means would be highest in the year students received special education services. Results revealed that when students’ EOG-M scores for all cohorts were combined, means were higher for the year they received special education services than they were the year before they received services. However, no difference was observed between means for the year students received services and the year after they received services. No support was found for Hypothesis 5, which asserted that EOG-R means would be higher in the year students received services than the year before and year after they received services. In addition, results indicated that cohort scores did not differ for the year before students received services (A1); between cohorts for the year they received special education services; or between cohorts depending on what grade that they received special education services.
Summary of Hypothesis Testing

When considered together, these results reveal support for five of the six hypotheses presented in Chapter Three. For the cohort comparisons used to investigate Hypotheses 1 and 2, special education students’ EOG-R scores improved (relative to their peers) when the cohort static method was used to identify group affiliation, but remained consistent when the cohort dynamic method was used. For Hypotheses 3 and 4, special education students EOG-M scores remained consistent compared to their peers, but they fell further behind their peers when the cohort dynamic method was used. The current study did not find evidence to support Hypothesis 5, which stated that students’ academic gains as measured by the EOG-R would be greater the year they received special education services than the year before they received services and the year after they were in special education. Finally, Hypothesis 6 stated that students’ academic gains as measured by the EOG-M would be greater the year they received special education services than the year before they received services and the year after they were in special education. Results revealed a significant effect for time, and post hoc analysis revealed that EOG-M scores were greater the year students were in special education than the year before they entered special education. However, no significant difference was found between the year students received special education services and the year they exited special education.
CHAPTER SIX

Discussion

The right of opportunity and access to education for students with disabilities has been well established through Supreme Court rulings, federal legislation, and more generally by the ethical standards of educators. Despite the awareness that students with disabilities should have the right to an appropriate education, educators and researchers have grappled with how to best serve children with disabilities. That is, the debate over special education is no longer one of access to education, but rather effectiveness of special education services. Indeed, for over 40 years, critics of special education have argued that special education is ineffective, citing a lack of empirical evidence to support the placement of students in segregated classrooms (Dunn, 1968), that placement in special education had an overall negative effect on student performance (Carlberg & Kavale, 1980), and that the identified achievement gap between special and regular education students grows larger as students get older (Trimble, 1998; Ysseldyke et al., 1998).

With the recent reauthorization of the Individuals with Disabilities Education Improvement Act (2004) and the 2001 authorization of the No Child Left Behind Act, students with disabilities are now required to participate in large-scale testing. Consequently, researchers have a new tool for measuring academic outcomes for students with disabilities, and addressing the question of special education effectiveness. The present study made use of this new tool to examine the effectiveness of special education by: (a) replicating and extending the results of one of the first studies to use large scale testing results to examine the achievement gap between students in general and special education, and (b) conducting a within-student comparison of achievement gains in general and special education.
The first five chapters reviewed previous research and legislation regarding the education of children with disabilities, described the hypotheses for the current study, provided details regarding methodology of the current research, and provided results of the investigation. Chapter Six will begin with a discussion of the current study's results as they are related to the two major aims of the current study, the research hypotheses related to those aims, and previous research. Following the two sections discussing results related to the achievement gap and within-student comparisons of general and special education outcomes, a general discussion will explore the significance of the current study in terms of special education effectiveness research. Broad implications of the study results for the field will be provided. Finally, limitations of the current study will be provided, as well as a discussion of future research needs.

Achievement Gap Comparisons

Summary of Findings (Hypotheses 1-4)

Ysseldyke and Bielinski (2002) conducted one of the first studies to use large-scale achievement test data to assess the effectiveness of special education. Their examination utilized student scores on the Texas Assessment of Academic Skills, and created effect sizes to describe the performance of special education students across grades four to eight, with general education students used as the reference group. Results revealed that the method used to identify student classification affected the magnitude of the achievement gap between special and general education students. When student classification remained stable over time, the effect size for the achievement gap remained stable. However, when students' special education status changed as their school special education classification changed (the
traditional way test outcomes have been reported), the achievement gap between special
education and general education grew larger as the students progressed through school. This
research was important because it showed that by reclassifying students every year according
to their school classification, special education was not credited for the effect it had on
student performance, as the highest achieving students tended to be the ones most likely to be
exited from special education services.

The first aim of the current study was to replicate the analyses conducted by
Ysseldyke and Bielinski (2002), but with large-scale test data from North Carolina. Similar
to Ysseldyke and Bielinski, it was expected that the achievement gap between special and
regular education in both reading and mathematics would be greater when the Cohort
Dynamic method was used to calculate effect sizes than when the Cohort Static method was
used. All hypotheses were confirmed. In each comparison (across subject area and cohort),
the effect size comparing students in special education to those in general education differed
depending on what method was used to determine group membership, and the achievement
gap effect size was always larger when the Cohort Dynamic method was used.

Although the current study’s results were similar to the findings of Ysseldyke and
Bielinski (2002) with regard to the comparison between methods of calculating the
achievement gap, they differed in terms of the trend in the size of the achievement gap across
grades. Ysseldyke and Bielinski found that the achievement gap for special education
students stabilized rather than grew larger with the Cohort Static method. The current study
results showed the achievement gap in reading (but not mathematics) decreasing over grades.
For mathematics, the effect sizes at each grade did not differ, suggesting the achievement gap
remained stable across grades.
Discussion of Achievement Gap Findings

The current study’s results strengthen the external validity of Ysseldyke and Bielinski’s (2002) research by replicating their findings using data from another state. However, they also reveal a continued troubling reality for special education students. Using both the EOG-R and EOG-M data and the Cohort Static method of reporting, students in special education were still performing nearly a full standard deviation below their regular education peers in eighth grade (EOG-R = -.85; EOG-M = -.86). Perhaps it is encouraging the achievement gap narrowed when using the EOG-R scores. However, the gap only narrowed .15 standard deviations over four years. If that trend continued, it would take 22 years for the special education students to catch up to their regular education peers. Although these data contradict the contention that special education students are falling further behind their peers as many special education critics have suggested (e.g., Gartner & Lipsky, 1989), they do not suggest that special education is “catching students up” to their regular education peers. As will be presented in the general discussion, these data suggest that closing the achievement gap completely, a stated goal of current national education policy, may not be realistic unless there are marked advances in the efficacy of treatments available within special education.

It is important to note the starting point differences between the sample in the Ysseldyke and Bielinski (2002) study and the current sample. Special education students in the first year of the Ysseldyke and Bielinski study performed approximately .5 standard deviations below their regular education peers. In the current sample, the special education students as a group scored between .88 (EOG-M) and 1.01 (EOG-R) standard deviation units below students in general education. One potential reason for the difference in size of the
achievement gap between the two studies, and an encouraging aspect of the current results, is the increased representativeness of the special education sample compared to the Ysseldyke and Bielinski study. In their study, only 60% of fourth grade special education students in Texas took part in the high stakes assessments, and by 8th grade this number dropped to 39%. Ysseldyke and Bielinski noted that those most likely excluded from testing were students with the highest needs and therefore likely to be the lowest performing special education students. For the current study, at least 85% of special education students took the EOG tests in each of the years, supporting Ysseldyke and Bielinski’s assertion that a more representative sample of special education students who partake in high stakes assessments would result in lower test scores for special education students as a group.

_Different trend line for the achievement gap as measured by the EOG-R._ The current study provided evidence that neither the reading or math achievement gaps grew over time when using the Cohort Static method to compare achievement of students in special education to those in general education. In fact, the trend line for the EOG-R with the Cohort Static method indicated the achievement gap grew progressively smaller from 4th to 6th to 8th grade.

A potential explanation for decrease in the achievement gap as measured by the EOG-R is the nature of intervention for reading in schools. Reading is a core skill that once learned can continue to improve over time (Torgesen et al., 2001). That is, once specific literacy skills are learned, students’ reading ability can continue to improve over time even when the specific reading interventions have been removed by continuing to practice the skills that they learned through intervention on their own and with the help of the general education curriculum. Effective math intervention, on the other hand, requires continuous
introduction of content so that children can continue to learn more complicated math skills (Maccini, Mulcahy, & Wilson, 2007). Therefore, given that the current study continued to include students in the ‘special education group’ even if services were removed after fourth grade, it is possible that the decreasing achievement gap as measured by the EOG-R was the result of student continued improvement in reading even if special education services were removed.

Policy implications for subgroup identification. Although the use of high-stakes assessments as a sole measure of student achievement is a highly contested practice (Wang, Becket, & Brown, 2006), the reality is that current educational accountability systems (e.g., NCLB, 2001) put a priority on student performance on these tests, and students and schools face consequences for poor scores. Special education students are increasingly being included in such systems, and their scores are treated in a similar manner as other demographic groups (Linn, 2005). That is, special education students, and by extension the schools and staff that serve them, are being held accountable for their performance on high stakes assessments, and this practice gives researchers and policy makers the ability to have a common outcome measure from which to judge the effectiveness of special education.

Traditional strategies for comparing the performance of students with disabilities to students without disabilities involve reclassifying students each year according to their special education status. As first proposed by Ysseldyke and Bielinski (2002) and replicated in the current study, this method of reporting results in a condensed group of low achieving students. However from a policy standpoint, where schools are not only judged by the overall school performance but also on the performance of several subgroups (NCLB, 2001), reclassifying student affiliation each year makes it easier to identify specific groups of
students who are struggling in a given school. In order for a school to be determined as having made adequate yearly progress, every identified subgroup (including special education students) must reach progress goals independent of the other groups in that school. Many states report that over 70% of special education students fall below expectations in reading and math on statewide achievement tests (Chudowski & Chudowski, 2009). Given that a recent three-state study found that the most frequent reason that schools fail to make adequate yearly progress is the performance of the special education subgroup (Eckes & Swando, 2009), the performance of this group is of particular importance to policy makers and schools alike. A reasonable question based on the struggle of this particular subgroup is, “How do we increase the performance of this subgroup of students?” Given the results of this study, perhaps another question should be, “Is the way this group of students is defined appropriate?”

When defining the groups for the current study using the Cohort Dynamic approach (the approach used for defining groups when states calculate adequate yearly progress), the students in the special education group were served by special education in every year. When the Cohort Static approach was used, the students were served by special education in the first year, but then may or may not have been served by special education in subsequent years. In other words, the group of students who were identified as having a disability that qualified them for special education in 4th grade may have only received services for that particular year. In subsequent years, they may have been exited out of the special education program and received their educational services from general education. If students who exited special education were to remain in the special education subgroup for the purpose of calculating adequate yearly progress, the problem of special education becoming a condensed
group of low achieving students would be addressed because special education as a service would continue to “get credit” for the progress made by students who were served by special education at some point. On the other hand, continuing to add students to the special education subgroup without removing students based on their current special education status would make judgments about the progress of this subgroup less precise. For example, if students are kept in the special education subgroup for the rest of their educational career regardless of their current special education status, a student identified as in need of special education services in kindergarten could be exited from special education in first grade, yet still be considered a special education student in 8th grade when determining the subgroups adequate yearly progress. It could be argued that the services the child received in kindergarten set him or her up for success for the rest of their academic career (Barnett, 1995). However, it could also be argued that it is inappropriate for a child in middle school who has not received special education services in seven years to be grouped in with students who continue to receive services for the entirety of their education career. In terms of educational policy where subgroup performance is calculated each year based on the special education status of each student for that year (NCLB, 2001), it seems unlikely that policy makers would seek to include students who did not receive services in the year of testing into the special education subgroup.

One of the reasons that group affiliation is so important is that special education is considered a subgroup when determining adequate yearly progress similar to race and socio-economic class. Current educational policy reflects the expectation that the general education-special education achievement gap will close at the same pace as other subgroups. However, a review of the special education literature reveals that special education students
start off far below general education students in terms of academic performance. Further, the special education subgroup is the only subgroup where limitations on ability to learn is a criterion for inclusion in the group (Eckes & Swando, 2009). These factors call into question the appropriateness of efforts to completely close the general education-special education achievement gap. This topic will be discussed further in the general discussion, where alternative approaches for monitoring special education students’ progress will be presented.

Within Subjects, ABA Hypotheses

A within subjects design was utilized to test Hypotheses 5 and 6. Hypothesis 5 predicted that in the group of students who were classified as general education students one year, special education students the following year, and then exited special education in the following year, EOG-R scores would be higher the year they received special education than either the year before or the year after they received services. Similarly, Hypothesis 6 used the same design and predicted that EOG-M scores would be higher the year students received special education than either the year before or the year after they received services.

Hypothesis Five

As stated in the Results section, this hypothesis was not supported. Student annual gains in achievement in reading did not change with changes in special education classification. One explanation for the failure to support Hypothesis Five is the age of the participants in the current study. The North Carolina EOGs are not given to students until third grade. By that time, many students with early reading difficulties have been identified and have received intervention. Often, early phonics intervention is very effective at improving students’ reading comprehension (Vaughn et al., 2009). Therefore, an analysis of
special education performance that does not take into account student gains in early elementary may not reflect the actual impact that special education has had on students’ reading performance. Because the current sample used EOG scores from North Carolina as the primary outcome measure, and these tests are not given until third grade, any gains made by students before third grade would not be reflected in any analyses of EOG test scores.

A second explanation for the failure of the present study to support Hypothesis 5 is that the EOGs were not an appropriate outcome measure for the current sample. High stakes assessments are expected to provide broad overviews of student performance in specific academic areas (Shriberg & Kruger, 2007). A hallmark of special education is individualized education strategies where specific outcomes are expected over a period of time. Without access to each participant’s IEP, it is impossible to say whether or not reading was an area of targeted concern for this group of special education students. Post-hoc analyses revealed that the achievement gap for this group of special education students and their regular education peers as measured by the EOG-R in the year they were in special education was only -.11, which suggests this group of students performed much better than the entire population of special education students in North Carolina. As discussed earlier, the achievement gap between special education and regular education students on the EOG-R when all special education students were included in effect size calculations (instead of only students who both entered and exited special education) ranged from -.94 to -1.01. Based on their performance on the EOG-R tests, it is possible that the students included in the ABA within-subject comparisons in the present study did not have reading goals on their IEP. If so, expecting these students’ EOG-R scores to improve in the absence of targeted reading intervention may have been unreasonable.
Finally, a potential explanation for the failure to support Hypothesis 5 is simply that special education does not improve students’ achievement in reading. However, as stated in Chapter Two, the question of special education effectiveness is so broad that it seems impossible to make generalized statements about the overall effectiveness of special education. Rather, the question of special education effectiveness should be limited to the participants in a study, the way that the participants were identified as special education students, and the time frame that the study took place in. Following this model, the current results would suggest that for students who transitioned into and out of special education between 4th and 6th grades or 5th and 7th grades, and who entered either 4th or 5th grade in either 2000 or 2001, special education was not effective for increasing the EOG-R scores. Although this narrow interpretation of study results precludes drawing any general conclusions about special education effectiveness, the future research section will discuss the need for continued investigations of special education outcomes using several different and specific operational definitions that precisely identify the sample and dosage of special education received, and then ascertain what outcomes are obtained under those conditions.

Hypothesis Six

Hypothesis 6 predicted that students’ EOG-M scores would be higher the year they received special education than either the year before or the year after they received services. This hypothesis was partially supported, as students’ EOG-M scores were higher the year they received services than the year before they received services, but there was no difference between the year they received services and the year after they exited special education. That is, students performed better on the EOG-M the year they entered special
education, but did not return to baseline the year the students were exited from special education.

The results from the current study are consistent with those found by Hanushek, Kain, and Rivkin (1998), who found that entering special education programs was associated with a gain of .17 standard deviations in mathematics when compared with students who remained in general education or special education programs. In a similar study, Ewing (2009) controlled for regression toward the mean and found that entering special education programs was associated with a gain of .18 standard deviations in mathematics. However, it was expected that EOG-M scores in the current study would return to baseline after students exited special education. Fuchs and colleagues (1996) found that one year after reintegration into general education about one half of students were never fully reintegrated or were moved to general education temporarily and then returned to special education. The main reason given for the high return rate to special education for participants in that study was that math achievement slowed considerably when students were placed in general education. Given that students who struggle in math typically require ongoing intervention (Maccini, Mulcahy, & Wilson, 2007) and the results of Fuchs and colleagues (1996) who identified poor math achievement as a main reason for return to special education at one-year follow up, it is surprising that the current study failed to fully provide support for Hypothesis 6.

One potential explanation for the partial support of Hypothesis 6 (i.e., EOG-M scores increased compared to the previous year when they began special education services) and the lack of support for Hypothesis 5 is the age of the participants involved in the current study and when they likely would have received special education services for math and reading. Early interventions (defined as interventions that take place before second grade) seem to be
more readily available in literacy than mathematics. An examination of the U.S. Department of Education’s “What Works Clearinghouse” website (www.wwc.gov) reveals that there are 171 early literacy interventions that have been reviewed, but only 73 early math interventions reviewed. Of these, there are 24 reading interventions that have research indicating positive or potentially positive results, and there are only two math interventions that have positive or potentially positive results. The number of effective early reading interventions compared to early math interventions underscores the priority that early literacy skills take in schools. Given the greater number of early intervention reading programs and the priority that early intervention seems to have in schools, it is possible that reading interventions are more likely to be the first targeted interventions given to special education students, with more targeted math interventions coming later in the elementary years. Therefore, students in the current sample may have been receiving math services for the first time in later elementary years (i.e., the years they took the EOG-R and EOG-M tests) despite having received early reading intervention during lower grades in their educational career. If that was the case, the current study would not have been able to detect gains made in reading when reading interventions were first introduced.

It is also a possibility that math goals were a part of the IEP for many of the students included in this sample, and therefore the special education services they were receiving specifically targeted math skills. The achievement gap as measured by the EOG-M for students in this sample was -.93 for the year before they received special education services. This achievement gap is more in line with the one observed for the special education population as a whole than was this group’s EOG-R scores for the same period (-.11). If the students received services directly targeting math and not reading, it is reasonable that math
scores improved the year the students received services while the reading scores did not improve.

A potential explanation for why a return to baseline was not found in the current study is the way that special education students were defined in the analyses. To be considered a special education student in North Carolina for purposes of statewide accountability, students have to be receiving special education services at the time of the EOG tests. Unfortunately, this definition could have resulted in a wide range of services that students could have received and still have been eligible for inclusion in the current study as a “special education student.” Given that the EOG tests are given in May each year, a student would have had to be exited from special education by May in order to again be considered a general education student. That means a student could fit the criteria for special education in the current study if they were exited from special education in September, or could also fit the criteria if they were exited in April. Despite the potential seven-month difference, all of these students were considered as not having received services that year in the present study. Fuchs and colleagues’ (1996) follow-up study of students who exited special education in mathematics took place one full year after students were reintegrated into general education. The students in the current study had been reintegrated for not more than eight months. Perhaps students’ achievement gains would have returned to “baseline” levels if achievement in general education had been assessed in a subsequent year.

In sum, the analyses related to Hypotheses 5 and 6 provided mixed results about whether the provision of special education services is associated with increased gains in achievement relative to gains observed in general education. It is important to note, however, that these results are equally important when addressing the question of special education. In
the general discussion to follow, it will be argued that special education research is similar to putting together pieces of a large puzzle. In truth, it is more complicated than that analogy and is more like trying to put one puzzle together when you are presented with the pieces of many different puzzles. When taking on that task, it would be important to identify the pieces that do not fit the actual puzzle you are putting together. For special education effectiveness research, one must not only find the pieces of the puzzle that fit, but also those that do not fit. The pieces that do not fit should be acknowledged and removed so as not to interfere further with the pieces that do fit. So although one hypothesis was not confirmed (Hypothesis 5) and another was only partially confirmed (Hypothesis 6), it is important for researchers to consider and report both significant and non-significant findings of special education effectiveness research so that they may better guide future practice.

General Discussion

The results of the present study illustrate that the method of reporting outcomes for special education students has an impact on the conclusions that are drawn about the effectiveness of special education. The results also provide some support for the beneficial role of special education on achievement, given that entrance into special education was associated with increased achievement gains in mathematics. Yet the biggest contribution of the current study may be as an illustration of how the wealth of data now available to researchers as a result of high stakes testing can be used to address questions of special education effectiveness. The current trend in school reform where school-wide high stakes assessments are the norm, is still in an infancy stage, and researchers are just scratching the surface of what analyses are possible given the data now available. This general discussion will examine the significance of the current study to the field of special education, how the
current study addressed limitations in special education effectiveness research, and the continued difficulty special education effectiveness researchers face given the enormity of the question. Finally, a discussion of reasonable standards for measuring special education students’ academic progress will be presented with suggestions for future research.

**Addressing Limitations of Special Education Effectiveness Research**

The question of special education effectiveness is enormous. By extending what was first observed by Leinhardt and Pallay (1982), who suggested that it was not the fact that students were in special education that was important but rather what happens as part of the services they receive, researchers can understand that global attributions of special education effectiveness are often inappropriate and misleading. Such is the case for researchers who suggest that the achievement gap between special and regular education students increases over time without considering students who move in and out of special education. These types of global attributions require additional research, such as the first four hypotheses of the current project, to refute those claims by using different, if not more precise, procedures for identifying special and regular education samples.

Unfortunately, the refutation of inaccurate claims by using equally global attributions does little to further the field of special education. After all, the end goal should be to increase the effectiveness of special education services so that students with disabilities have more success. This goal can be accomplished by providing evidence to frontline practitioners in the field about what parts of special education work, for whom, and in what timeframe. That is, a continued debate about the overall effectiveness of special education seems useful for political purposes given the cost of special education, but does little to improve the service delivery of special education.
Special Education Effectiveness

Despite the thousands of research articles that have been published over the years, it seems that special education is still not operationally defined well for research questions, and therefore not well understood. While some researchers continue to address the question of special education effectiveness in global terms, others see the question of special education effectiveness as a giant puzzle (e.g., Forness & Kavale, 1994). Just as a piece of a puzzle means very little until it is combined with other pieces to make a coherent picture, the individual studies of special education effectiveness provide very little in isolation. Therefore, some researchers have conducted meta-analyses of special education effectiveness research findings, and Forness (2001) expanded this idea and conducted a mega-analysis (a meta-analysis of meta-analyses). Beyond the typical methodological shortcomings of meta-analyses, meta-analyses in special education effectiveness research continue to be faced with the same difficulties that individual studies have in this field: no consistent way of measuring progress and difficulties with identifying appropriate comparison groups.

Identification of appropriate outcome measures. A limitation of special education effectiveness research, as discussed in Chapter Two, has been the identification of appropriate outcome measures. The use of EOG scores in the current study was chosen to address this limitation, as high stakes tests are an accepted measure of student, and more broadly, school performance. Schools in North Carolina are held accountable for students’ performance on these tests, and it is expected that students make “adequate yearly progress” as measured by these exams. Each school has a report card where it is documented how well each school’s students performed on the EOG tests each year. Along with the whole school population, schools are also responsible and held accountable for the adequate yearly progress of several subgroups, and special education has been identified as one of these
subgroups. So in a very general way, schools are already being held accountable for special education effectiveness. However, this global approach to accountability is not sufficient to answer the decades old question of special education effectiveness for two reasons.

First, despite evidence that achievement gap did not grow over time, results of the current study revealed that students in special education did not “catch up” to the mean achievement scores of regular education students. This finding adds additional evidence to the body of research suggesting that students with disabilities continue to lag behind their non-disabled peers (e.g., Schulte et al., 1998). Although demonstrating special education students’ year-by-year gains that are on par with the progress of the general education student population is encouraging, it is unlikely to satisfy special education critics who charge that special education is too expensive just to maintain the achievement discrepancy between special and regular education students (e.g., The Wall Street Journal, 1993). Thus, research must document that special education students perform better as a result of receiving special education services than they would if they did not receive such services.

Second, global evaluations of special education effectiveness such as those that can be inferred through school report cards are simply too broad to guide the field. Realistically, the global accountability of special education as a subcategory that is in use now only serves to report how schools’ special education students performed. Although these data can inform us about which schools are struggling, without more data, such as disaggregating the data by disability type, type of special education interventions provided, and the amount of time services were given, there is no way to tell why each school’s special education population did not make adequate yearly progress. Further, it is unlikely that all of the special education students in a school would not make adequate yearly progress in the same year. More
detailed analysis would allow for more categories of students, projections of adequate yearly progress to be made based on these categories, and a more meaningful evaluation of a school’s special education students’ performance that can guide future planning.

Operational definition of a ‘special education student.’ As discussed in Chapter Two, from the inception of special education through the 1980s, special education was widely regarded as a place rather than a set of specially designed services. Although special education is rarely regarded as a place anymore and is instead regarded as a continuum of services, remnants of this discrete thinking still haunt special education effectiveness research. The continuum that now helps define special education does not extend into other school-based services, meaning that the discrete, categorical nature of special education still exists. That is, within special education services there exists a continuum of services, however there are still discrete categories that separate special and regular education students. Categorical thinking presents numerous problems when researchers try to identify special education students. For example, if a researcher wants to collect data about the effectiveness of special education for 4th grade students based on EOG data, should students who received services from kindergarten through 3rd grade be considered special education students, even though they made the categorical jump out of special education and no longer received services in fourth grade? A hallmark of scientific investigation is the specific operational definition of the sample, and this is an area where special education effectiveness research has been lacking.

The current study addressed this methodological limitation in two ways. First, in a global analysis of the achievement gap between special and regular education, the current study provided evidence that when special education is given “credit” for contributing to the
Special Education Effectiveness

success of a student, the achievement gap does not continue to grow. Although this is an interesting finding and can be used to combat critics’ calls that the achievement gap continues to grow, it still may be too broad to truly guide practice. The analyses did not refine the global definitions of the student groups, and instead relied on school identified special education status for the years in question, which is the practice typically used to track the achievement gap over time. Unfortunately this type of global analysis does not provide evidence that special education was more effective for individual students than regular education would have been. Although we can infer that students were struggling academically in regular education and thus were identified as being in need of special education, this type of analysis does not allow a researcher to use data to answer the question of whether or not special education was more effective for individual students than regular education.

The second way that the current study addressed the issue of operationally defining the target groups in special education was to use a within subject design, and this methodology seems to hold more promise for guiding future research and practice. Specifically, using a within subjects design for students who transition in and out of special education allowed for a determination of whether students performed better in years they received services versus years they did not receive services. As first proposed by Ysseldyke and Bielinski (2002), one purpose of the current study was to carry out this type of investigation in order to scaffold future research using this methodology. The current study focused on the specific years that students transitioned in and out of special education, and this type of analysis lends itself to being operationally defined in more specific ways. If results of this study are considered in isolation, it is unlikely to have a major impact on the
field of special education effectiveness research. Therefore, the power in the current study is the evidence that a within subjects design using EOG test data can be used to answer parts of the question of special education effectiveness, and the more of these pieces that are put together (i.e., the more ways that the target group is operationally defined), the better equipped we will be to answering questions of special education effectiveness, and guiding frontline service providers about how and why special education is effective for students.

**Limitations**

When interpreting the results of the present study, there are several limitations that should be considered. Three important limitations will be elaborated here: use of a non-experimental research design, limited generalizability of results given the representativeness of the sample, and missing information.

*Research design.* The first limitation of the current study is its correlational research design. No direct manipulation of variables occurred and no direct causal link between special education placement and performance on EOG tests can be established. This is a typical limitation in special education research, as random assignment to special education and general education for students with disabilities presents legal and ethical issues.

*Generalizability of study results.* Great effort was put forth in previous sections to detail the importance of considering each special education effectiveness study as a piece of a giant puzzle. The current study is no exception. For the achievement gap analyses, students were placed into cohorts based on their status in 5th or 6th grade in specific years. The set of years selected was determined because there were EOG data that would allow for analyses over five consecutive years. Therefore, these results only truly reflect the effectiveness of
special education programs for students who at least received services in either 5<sup>th</sup> or 6<sup>th</sup> grade (depending on the cohort). As will be discussed later, the current data did not provide evidence of how long within a year a student received special education services. Further, the analysis did not include how many years before the targeted year a student received services. Despite efforts to operationally define the special education group, there is still the potential for very big differences in terms of the influence of special education on a student’s progress based on the amount of time a student was in special education. Therefore, a major contribution of the achievement gap analyses was to provide an example of how achievement gap data can be tracked across years in a way that does not make the special education students a concentrated group of low achievers (i.e., by using the Cohort Static Method), thus combatting critics’ arguments that the achievement gap continues to grow over time. To make more specific interpretations without repeating the analyses for different grades, disability type, and time spent in special education would be inappropriate.

Similarly, overgeneralization of the results of the ABA analyses using EOG-M data is a potential limitation of the current study. Given that the students in the sample were able to transition into and out of special education in one year’s time suggests that the disabilities in question were likely mild in that the students could be served without receiving special education services. Further, given that special education for students in the sample was delayed until 4<sup>th</sup> or 5<sup>th</sup> grade (depending on cohort), it seems unlikely that these students had severe disabilities that resulted in low functioning, as they would have been identified in earlier years and stayed in special education throughout the current study’s targeted years.

**Missing information.** There is information that was unavailable for the current study that would have been beneficial, especially for operationally defining the special education
groups in the ABA analyses. Specifically, there was no information available as to what services students received while they were in special education. As was discussed earlier, special education is a set of services that can vary widely, making this information very valuable in an evaluation of special education effectiveness. Without this information it is possible that the children included in the analyses could have been, and likely were, receiving very different services. As such, the information gathered lends itself to more global evaluations of special education effectiveness rather than identifying “what worked” in special education and for whom. Similarly, there was no information provided about the time frame for which students received services. To be identified as a special education student in a given year requires that for at least part of the school year a child received services. For some children this could have meant the whole year, and for others this could have meant weeks. Therefore, there may have been a dosage issue for not only the intensity of services received but also the amount of time they received them.

Finally, there was no information about IEP goals for individual students in the present sample. It is reasonable to expect special education to have an impact on students’ EOG-M scores if the student is receiving services to reach specific math goals, and it may not be reasonable to expect an increase in EOG-M scores if the child is not receiving specific math-based services. However, some students without math goals on their IEP may still be expected to have an increase in their EOG-M scores if IEP goals address other general academic topic areas, such as on-task behavior and work completion. Had each student’s IEP goals been available, correlational links could have been established between goals and EOG test scores. The next section will discuss the advantages of repeating the research design used in this study with other datasets for future research. If researchers had IEP information
available to them when they replicate the methodology of the current study, a growing number of correlational links could be established between EOG scores and IEP goals, dosage of special education (i.e. the number of years needed to benefit based on student characteristics), specific services received, and a number of other characteristics. Combined, these data could have a large impact on the ability to guide frontline special education teachers and decision makers.

**Directions for Future Research**

The current study found support for the effectiveness of special education, and perhaps more importantly identified and carried out a promising research design that can be used to continue this line of research. This section outlines suggestions for future research and implications for practice based on the findings and limitations of the current study.

**Research design.** The results of the current study suggest that a causal relationship may exist between late elementary students’ performance on EOGs and their special education status for that year. However, given the correlational design of the present investigation, no such conclusions can be drawn. Although unethical to randomly assign students to special or regular education, future research could seek to randomly assign groups to what services they receive once identified as needing special education services. Further, there is little in the research literature that suggests special education teachers document the level to which they follow through with intervention and specialty instruction practices. By manipulating classroom variables (e.g., type of instruction/intervention used, length of intervention, pupil-teacher ratio), randomly assigning students to treatment condition based on these factors, and having a measure of fidelity to document that the services were
implemented as intended, researchers could potentially identify a causal link between aspects of special education and achievement.

_Similar research design with different operational definitions._ All states now use high stakes assessments, and although these assessments should not be considered equal, they are all similar in that students and schools are held accountable based on student performance. As such, researchers now have a wealth of data from every state that can be used to analyze student progress. If these data are used to evaluate special education effectiveness, one of the major limitations of special education effectiveness research will be addressed. Further, as identified in this study, researchers can use a within subject design to address the issue of an appropriate comparison group. Therefore, if methodology utilized in the current study is consistently used in the future to address the many subsets of special education students and time frames that can be identified, there exists the potential for research to appear more like a puzzle that fits nicely together rather than an incoherent, disjointed set of individual studies.

_See alternative measurements for younger students._ One of the potential reasons given for the failure to find support for Hypothesis 5 in the current study was that older elementary students may not benefit from special education in terms of improvement in EOG-R scores in the same way as younger elementary students. Unfortunately, high stakes assessments in most states do not begin until 2\textsuperscript{nd} or 3\textsuperscript{rd} grade. Future research should seek to find generally agreed upon assessment techniques that can be used uniformly across research studies. The reason that high stakes assessments are an attractive option for researchers is that current educational policy has made high stakes assessments a universally acknowledged option for assessing student growth (NCLB, 2001). To capitalize on this advantage that researchers now have, future research should seek to push the agenda of an accepted
measurement of growth for younger students. Although there are many criticisms against the use of high stakes assessment (Wang, Becket, & Brown, 2006), the possibility of more child-data that could serve to guide schools in how to better serve their students is certainly something that should be investigated further. In lieu of more consistent, global assessments for younger students, special education effectiveness research for younger students will likely continue as a disjointed endeavor.

The current study identified a very specific subset of children and analyzed their EOG scores for a specific set of time. Future research should look to build on the findings of this research and methodology by changing the operational definition of the sample and time frames for which they are followed. For example, the current study used a within subjects design and considered dosage by limiting the number of years a student received special education to one year. Future research could vary the number of years that a student received special education and use the same within subjects design and EOG data. The two studies could then be compared to see if the number of years a student received special education had an effect on EOG performance. Further, more information could be gathered about what services were offered to the children, and the same methodology could be used to compare the effectiveness of different interventions. By using a widely accepted method of reporting student progress and within subjects design, the possibilities for comparison seem endless.

**Implications for Practice**

*Reporting of student progress.* Current educational policy uses high stakes assessments to hold schools accountable for their students’ performance. In addition to being held accountable for the entire student population of their school, schools are also held
accountable for several subgroups in isolation, including special education students (NCLB, 2001). As was shown in the current study, different methods of reporting student affiliation can have a dramatic effect on trends and comparisons between groups. The current study provided evidence that when students’ special education affiliation is redefined each year, the achievement gap between special and regular education gets bigger over time.

This nomothetic approach to accountability seems at odds with the individualized nature of special education. For the highest achievers in special education, the academic goals they have in place may serve to springboard them out of special education. In practice, having the highest achievers exit special education and the lowest achievers enter special education can have a negative impact on a school’s report card. Without data on students who were reclassified being considered by state education agencies, not only may schools not receive the credit they deserve for helping high achieving special education students reach their goals, they may face a form of double jeopardy because they are going to have lower achievers enter special education each year. If current policies of identifying academic yearly progress (AYP) continue, state education agencies should consider reclassification data and use cohort static methods as part of their consideration of a subgroup’s progress.

*Increased access to data for decision makers.* The methodology outlined in this study holds promise for allowing researchers to provide school staff with relevant information about outcomes for children who receive special education services, both at the school and individual student levels. As mentioned earlier, this study in isolation only provides information about a small subset of children. To be useful to all schools, this research methodology would need to be repeated many times over using different operational definitions for special education students and services.
By combining research data that used similar methodology, the research questions could become more targeted and the results more globally generalized over time. With the acknowledgement that no one study would answer the question of special education effectiveness and shifting attention to answering pieces of the question that others haven’t asked, the net result will be a more general understanding of what interventions work for what children and in what time frame. If the research were compiled into a coherent database, school staff could refer to the database when considering what services would benefit a particular child and the time frame that progress could be expected. With the availability of empirical data, there is reason for optimism for establishing large-scale databases reflecting typical progress of students with a variety of disabilities (Stone & Doane, 2001). Thus, the line of research methodology presented in this study has the potential to serve as a guideline for schools to make placement decisions.

Realistic expectations for special education students. Although future research should continue to identify more precise ways to measure special education effectiveness, national education policy is likely to continue to require more global assessments of student progress. The NCLB Act (2001) requires that by the year 2014, 100% of students must reach the proficient level or above in math and reading. However, based on the current definition of proficient (or adequate) yearly progress (AYP), many schools will fail to meet the NCLB requirement by 2014 (Linn, 2003). In the 2008-2009 school year, only 67% of schools in the United States made AYP (National Alliance for Public Charter Schools, 2011). Schools are required to report the results of students with disabilities, limited English proficient students, and economically disadvantaged students separately. Although requiring schools to report these data separately provides a mechanism for monitoring the achievement of lower
performing groups, this requirement increases the number of ways that schools can fail to meet AYP because every subgroup is required to meet AYP (Linn, 2005). If one subgroup fails to meet AYP, the entire school is deemed to have failed.

One mechanism that has been established to ensure that schools get credit for outstanding improvement is the Safe Harbor provision (Linn, 2005). Under this provision, even if a subgroup within a school fails to meet AYP requirements, the school can still be considered as having met AYP if there is a 10% reduction in the number of students who fail to meet AYP within a given subgroup. However, this level of reduction may be unreasonable, as very few schools have had that level of success (Linn, 2003). Instead, Linn (2005) calls for Safe Harbor to apply for schools who have a 3% reduction in the number of students who fail to meet AYP within a given subgroup. However, it is unclear if even a 3% reduction is a reasonable goal.

In a review of high stakes assessment data for reading in Colorado, Betebenner (2008) found that none of the students who performed one standard deviation below the mean in 6th grade were able to move into the proficient range in 7th grade. Further, only 30% of students who performed at the 10th percentile or below in reading in 6th grade were able to move into the partially proficient range in 7th grade. That is, a rate of growth falling at the 70th percentile (using growth norms based on the entire student population) was required for these students to move from the unsatisfactory to partially proficient range, and even with that amount of growth these students still did not meet the definition of proficient. Combined with data from the research literature (including results from the current study) that indicate students with disabilities tend to perform approximately one standard deviation below the mean on high stakes assessments, results from Betebenner’s (2008) analysis suggests that
even a 3% decline in the number of students with disabilities who fail to meet AYP may be unreasonable.

Betebenner (2008) called for the establishment of achievement growth reference charts as a way of establishing AYP benchmarks. Rather than using fixed targets that are not data driven, such as the NCLB Act (2001) calling for 100% of students making AYP, Betebenner (2008) outlined a model where student achievement could be monitored in much the same way that pediatricians use height and weight charts to explain the physical growth of a child relative to other children. The height and weight charts were developed through the analysis of thousands and thousands of children so that percentiles could be established. If a child is found to be at the 20th percentile for height in one year, that child is deemed to be growing appropriately if they continue to be at the 20th percentile the next year. With the availability of statewide assessment data, Betebenner suggests that achievement growth charts, which he referred to as student growth percentiles, could be established in much the same way. These data would allow states to use data to determine typical, below average, and above average growth, as well as set expected growth goals.

Although promising in that AYP goals could be data driven rather than formulated with arbitrary cutoff points, Betebenner’s (2008) proposal for student growth percentiles may still be too global to accurately predict the performance of special education students. As discussed above, schools are required to disaggregate data so that the performance of many subgroups, including students with disabilities, is reported separately. However, Betebenner proposes that students who perform in the same decile should be grouped together for the purpose of creating growth percentiles, regardless of their subgroup status (e.g. students with and without disabilities who perform in the first decile would be grouped together). Given
that students with disabilities historically perform consistently below their non-disabled peers, having their performance compared to the performance of students who perform in the same decile regardless of their special education status may be inappropriate. That is, rather than only using decile scores as a way of grouping students and developing expected performance criteria, perhaps students should be grouped according to decile scores and subgroup (e.g. special education status). These percentiles could be established in the same way that the proposed growth percentiles would be established, but would be more sensitive to the expected growth for a given subgroup population.

Unfortunately, special education is quite different from many of the other subgroups identified in NCLB (2001) in that the population of special education students is less stable. The features that define other disaggregated subgroups, such as race and socio-economic status, are unlikely to change from year to year. Therefore, tracking student progress across years is much more feasible for other subgroups, as there are fewer questions about what group they should belong in on a year-to-year basis. Special education is dissimilar from these groups in that students move into and out of special education frequently, with some students having multiple transitions over their academic careers. If growth percentiles were to be established for special education students, a consistent mechanism would have to be established for whether or not to include students who were exited from special education in the special education subgroup in subsequent years. If a consistent method of identifying students in the special education subgroup, the problem of special education becoming a more condensed group of low achievers would be accounted for. Even if students were reclassified each year (i.e., the Cohort Dynamic Method), growth expectations using percentiles should still be appropriate because the percentiles would have been derived from
Special Education Effectiveness

past years’ data, which would have included a condensed group of low achieving students in the special education group of older students. Having a consistently defined special education group would allow for the development of special education growth percentiles, which would allow for an appropriate comparison group for which future special education students could be compared, thus lessening the dependence on closing the achievement gap and instead focusing on reasonable, data driven goals.

In summary, the present study aimed to provide evidence that when using a cohort static method of reporting student affiliation, the achievement gap between special and regular education does not increase over time. In addition, the current study utilized a within subject design to investigate whether or not students’ performance on high stakes assessment improved as a result of receiving special education services. The current study replicates and expands upon previous research and provides an example of how a within subject design can be used to further address the question of special education effectiveness. Finally, the current study provides additional evidence to the growing body of research that suggests students’ high stakes assessment scores are useful for evaluating the effectiveness of special education.
REFERENCES


Raleigh, NC: Public Schools of North Carolina, North Carolina Department of Public Instruction, Raleigh Division of Accountability/Testing.


