DOES HOW STUDENTS ARE ASSIGNED TO CLASSROOMS MATTER? AN EXAMINATION OF RELATIVE ACHIEVEMENT IN TRACKED AND UNTRACKED MIDDLE GRADES LANGUAGE ARTS CLASSROOMS.

Christine M. Fierro

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Education in the School of Education.

Chapel Hill
2014

Approved by:
Fenwick W. English
Lora Cohen-Vogel
James Veitch
ABSTRACT

Christine M. Fierro: Does How Students are Assigned to Classrooms Matter? An Examination of Relative Achievement in Tracked and Untracked Middle Grades Language Arts Classrooms (Under the direction of Fenwick W. English)

Even with the controversial history of tracking students by ability and its possible differential, socially reproductive effects on student outcomes, tracking remains a common practice in public secondary schools. The purpose of this quantitative study was to investigate the relationship between students’ performance on state standardized tests and the type of classroom assignment practice employed. Specifically, students were tracked by ability for English Language Arts (ELA) two consecutive years. Their average performance was compared to the next year’s result when, at the same school, the same students were instructed in mixed ability ELA classrooms. With persistent achievement and resource gaps, continued pressures of high stakes testing, and the recent advent of including student performance data in educator evaluations, it was both timely and relevant to re-examine student to classroom assignment practices and their relationship with student achievement.

Taking advantage of a unique site in which most students experienced both “treatments” of tracked and mixed ELA instruction, changes in student performance were more attributable to time-varying factors, such as the type of classroom assignment, as opposed to time-invariant characteristics, like race, gender, or ability. Multilevel modeling accounted for the nesting of students within classrooms, while other factors such as teacher sequence, race, sex, and initial ability were also included in the model. Overall, non-advanced students who were mixed by ability with advanced students had the most significant achievement gains. Other groups also had
gains, though not to a statistically significant level. This finding, with replication, offers promise for the narrowing of the achievement gap between advanced and non-advanced students. As this gap mirrors racial and socioeconomic lines, also seen in this study, mixed ability classrooms may lead to more equitable outcomes, thereby also affecting future life conditions. Educational leaders must be cognizant of how and why student to classroom decisions are being made, paying attention to both results and antecedents. Similarly, as teachers play a critical role in student achievement progress, also supported by this study, leaders must develop and support teachers so they can best meet the varying needs of students.
To my parents, Sally and Lou, to Michael, and to Tyler. Thank you for your love, support, and never-ending belief in me in every facet of my life. I love you.
ACKNOWLEDGEMENTS

I want to thank several groups of people who all played important roles throughout my work toward this Educational Leadership doctorate degree. First, I want to thank all of my colleagues with whom I work on a daily basis. Your belief in my ability to continue to work full time and complete this degree always encouraged me to commit myself to both my job and my degree. Second, thank you to all of my classmates for your candor, insight, and overall contributions to my development. Knowing other educational leaders committed to social justice has been inspiring and has offered hope for a different future we can create.

Thank you to the University of North Carolina at Chapel Hill for your dedication to hiring top quality professors to serve your students. I also want to extend my gratitude to the adults and students who staff the Odum Institute—you are an invaluable resource for all of us. Thank you to all the professors with whom I have been able to work, both at North Carolina State University and at UNC Chapel Hill. You each contributed to my development as a student, a leader, and a person, and I am grateful for each of you.

Specifically, I would like to thank my committee for your time, support, guidance, and specific contributions through the dissertation process. Dr. Veitch, thank you for insisting on the maintenance of a practitioner perspective throughout my work and for reminding me of the theory-practice challenges. Dr. Cohen-Vogel, thank you for pushing me methodologically and for encouraging my conference proposal submissions. Lastly, a huge debt of thanks to my advisor, Dr. English—you helped me keep my head up when unforeseen obstacles arose,
encouraging me to keep going. Your reassurance and support greatly helped me to finish with a study of which I am proud. Thank you.

I hope to represent all of you and what I have learned well as I continue to work in service to public education.
TABLE OF CONTENTS

LIST OF TABLES ............................................................................................................. xiii

LIST OF FIGURES ........................................................................................................... xiv

CHAPTER ONE: STUDY OVERVIEW .............................................................................. 1

Introduction .................................................................................................................. 1

Tracking’s Troubling Historical Background ................................................................. 3

Problem Statement ........................................................................................................ 4

Delimitations and Context of the Study ......................................................................... 8

Purpose of the Study ..................................................................................................... 10

Major Research Hypothesis ......................................................................................... 10

Additional Research Hypotheses .................................................................................. 10

Rationale for the Study .................................................................................................. 10

Significance of the Research ......................................................................................... 12

Assumptions and Limitations of the Study ................................................................. 13

Definition of Terms ...................................................................................................... 14

Chapter Summary ....................................................................................................... 18

CHAPTER TWO: LITERATURE REVIEW ..................................................................... 20

Introduction .................................................................................................................. 20

The Problem .................................................................................................................. 21

The Historical Background of Grouping Practices ...................................................... 23
Rationale for Choice of Participants and Sample Size .......................................................... 81

**Research Design** ............................................................................................................. 83
Overall Design ..................................................................................................................... 83
Variables of Interest .......................................................................................................... 84
Other Variables .................................................................................................................. 85

**Procedures** ................................................................................................................... 85
Existing Data Sets .............................................................................................................. 86
Data Compatibility ............................................................................................................. 87

**Ethics, Validity, and Reliability** ..................................................................................... 87
Ethics ................................................................................................................................... 88
Validity ................................................................................................................................. 88
Reliability ............................................................................................................................... 90
Possible Threats to Reliability and Validity ...................................................................... 91

**Analysis and Statistical Procedures** ........................................................................... 93
Dependent and Independent Variables .............................................................................. 93
Initial Data Analysis ............................................................................................................ 94
Modeling the Data ............................................................................................................... 95
Building a Multilevel Model .............................................................................................. 96
Adding Predictors to the MLM ......................................................................................... 98
Significance Level ................................................................................................................ 101
Rationale for Statistical Procedures .................................................................................. 102

**Limitations and Significance** ....................................................................................... 103
Limitations ............................................................................................................................ 103
Significance .......................................................................................................................... 104

Chapter Summary .............................................................................................................. 104

CHAPTER FOUR: RESULTS ................................................................................................. 105

Introduction ......................................................................................................................... 105

Descriptive Statistics .......................................................................................................... 105

Performance Metrics .......................................................................................................... 108

Analysis via Multilevel Models .......................................................................................... 117

  Including Predictors in the MLM ....................................................................................... 121

  Interpreting the Final Multilevel Model ............................................................................. 136

  Rechecking the Assumption of Normally Distributed Residuals ..................................... 139

Summary ............................................................................................................................. 141

CHAPTER FIVE: DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS .................. 142

Introduction ......................................................................................................................... 142

  Major Research Hypothesis ............................................................................................... 142

  Additional Research Hypotheses ....................................................................................... 143

Interpretation of the Results ............................................................................................... 143

  Results Related to Sub-Hypothesis 1 ............................................................................... 144

  Results Related to Sub-Hypothesis 2 ............................................................................... 148

  Results Related to Sub-Hypothesis 3 ............................................................................... 149

  Results Related to Overarching Research Hypothesis ..................................................... 153

Implications of Results ........................................................................................................ 159

  Implications for School Districts ....................................................................................... 159

  Implications for Universities and Other Preparers of School Leaders ............................. 163
Limitations of the Study and Recommendations for Future Research .................. 166

Limitations ......................................................................................................................... 167

Recommendations for Future Research ............................................................................. 170

Conclusion ............................................................................................................................ 175

APPENDIX: NORMAL CURVE EQUIVALENTS ................................................................. 180

References .............................................................................................................................. 181
LIST OF TABLES

Table 1: Overall Cohort Descriptive Statistics ................................................................. 106
Table 2: Descriptive Statistics by 8th Grade Class Period ................................................ 107
Table 3: Descriptive Statistics by Grouping Sequence Trajectory ...................................... 107
Table 4: Proportion Minority by Class Period, Grades 7 and 8 ........................................ 108
Table 5: Performance Metrics by Race .............................................................................. 110
Table 6: Performance Metrics by Grouping Sequence ...................................................... 111
Table 7: Descriptive Performance Statistics by 8th Grade Class Period ............................. 115
Table 8: Analysis of Variance for the Change in NCE, 8-6/7avg, by Grouping Sequence .............................. 116
Table 9: Analysis of Covariance for the Change in NCE, 8-6/7avg, by Grouping Sequence, Conditioned by 8th Grade Teacher .......................................................... 116
Table 10: Results for Empty Means Random Intercept Model ........................................ 119
Table 11: Results for Adding in Grouping Sequence Predictor ........................................ 122
Table 12: Results for Adding in Teacher Sequence Predictor ......................................... 124
Table 13: Results for Race as a Predictor ........................................................................... 127
Table 14: Results for Including Sex as a Predictor ............................................................ 129
Table 15: Results for Including Grand Mean Centered Proportion of Class Male as a Predictor .............................................................. 131
Table 16: Results for Including Initial Ability as Predictors at Both Levels One and Two .............................................................. 133
Table 17: Results for Including Having a Support Class as a Predictor at Levels One and Two .................................................................................. 135
Table 18: Hypothetical Marginal Means Based on MLM .................................................. 155
Table 19: MLM-based Teacher Sequence Pairwise Comparisons .................................... 165
LIST OF FIGURES

Figure 1. The interaction of the guiding conceptual framework
with the study design. ........................................................................................................ 76

Figure 2. Change in ELA classroom assignment practices between
7th and 8th grades. ........................................................................................................ 81

Figure 3. Visual representation of 8-6/7avg NCE change scores
by grouping sequence. .................................................................................................... 112

Figure 4. Overall change in NCE by 8th grade class. ................................................ 120

Figure 5. The distribution of error terms/residuals by grouping sequence .................. 121

Figure 6. Distribution of residuals from the final MLM .............................................. 140

Figure 7. Normal Q-Q plot of the final MLM’s residuals ............................................. 141

Figure 8. Overall cohort distribution of African-American, White, 
and Latino students. ....................................................................................................... 150

Figure 9. Distribution of African-American, White, and Latino students
in advanced and non-advanced 7th grade classes. ..................................................... 151

Figure 10. Distribution of White, African-American, and Latino
students by 7th and 8th grade class period ................................................................. 151

Figure 11. Actual distribution of students by race and level by bars
versus equitable distribution by lines imposed on bars ............................................. 152

Figure 12. Graph comparing overall NCE change score trajectories by race .......... 153

Figure 13. NCE initial values and change score slopes by grouping sequence ......... 156

Figure 14. NCE change scores regressed on prior ability proxy
across 8th grade classes ................................................................................................. 158

Figure 15. NCE change scores versus prior ability proxy across
8th grade classes with reference lines ........................................................................ 159

Figure 16. 8th grade NCE gains by class period and grouping sequence ................ 162
CHAPTER ONE: STUDY OVERVIEW

Introduction

America is known as the land of opportunity. Similarly, American public schools have been proclaimed to be the great equalizer of opportunity for the students they serve. All students, regardless of their personal characteristics or family background, have access to a free and appropriate public education (EAHCA, 1975). Multiple times the court system has ruled in support of students’ rights to education, though it has been unwilling to go so far as to proclaim it a constitutional right (Lurie, 2013). Do American public schools truly provide an equitable educational experience to the students they serve?

Paradoxically, over time public schools have both changed greatly and largely remained the same. Remarkably, if adults of most any age were to walk into a school today, the experience would be familiar in both structure and feel (Bidwell, 2001; Lortie, 1975; Payne, 2008).
Simultaneously, since their origination, public schools have changed significantly in two ways: first, one-room schoolhouses gave way to multiple rooms and graded schools; second, school district consolidation occurred. In both cases, schools and school systems responded to larger and more diverse populations by adapting their organizational structures. Interestingly, while the United States continues to experience substantive changes in its demographic profile, the organizational structure of schooling has yet to respond with a third parallel alteration.

Absent overarching structural adjustments, multiple waves of change ensued within the school walls, including the age to which education was compulsory. As the Industrial Revolution
took hold, school curricula changed, as did the expected school outcome—no longer would an agrarian-focused eighth grade education sufficiently prepare students for success. Students were required to remain in school for additional years, and many students continued into upper grades. Another major change revolved around how students were assigned to classrooms for instruction. Following World War I, the intersection of an abundance of multiple-choice intelligence tests and an infatuation with efficiency emboldened the practice of assigning students to classrooms by ability, or tracking (Wraga, 1994). Indeed, school leaders used these first intelligence tests for the explicit purpose of having objective criterion on which to sort and select students into different levels of classes. Even though the resultant groups followed distinct racial and socioeconomic patterns, the application of the instrument persisted (Ansalone, 2003).

As a result, comprehensive high schools began to include vocational preparatory courses of study, or tracks, along with more academic, college-oriented tracks. Intelligence test results played a large role in determining into which track or course of study a student would enroll (Wraga, 1994). In line with effective business structures, educators believed they could more effectively and efficiently address students’ varying abilities and interests by creating more homogenous classes. Yet rational organizational theory tactics may not transfer so directly to schools, as Gamoran, Nystrand, Berends, and LePore (1995) asserted. The dynamic nature of the teaching-learning process challenged the rationality of schools as organizations (Cuban, 2004; Gamoran, et al., 1995). Furthermore, grouping students by ability allocated status and hierarchy mirroring students’ external family status, thereby raising questions about the democratic ideal of equal opportunity. Did tracking accomplish its purported goals of more efficient and effective instruction leading to improved student learning?
Tracking’s Troubling Historical Background

Grouping students into classrooms by ability, often referred to as tracking, has engendered a great deal of controversy. Early examples of how schools grouped students by perceived ability likely shaped the current practices and perceptions related to tracking. Within American schools, a form of tracking, or separating students into different instructional settings, can be traced back to the post-Civil War era when distinct groups of Southerners traveled North for greater life opportunities (Biafora & Ansalone, 2008). Similarly and in a more widespread manner, a greater number of immigrants arrived in America in search of a better life. Children from both migrant groups struggled to succeed in their new schools and were notably different from their receiving communities in both socioeconomic status and ethnicity. Schools responded by separating the new and different students into their own classes and courses of study. While insidious racism persists, much of the controversy about tracking today stems from an overall shift in society’s racial equity beliefs. These espoused beliefs of equity appear to have not translated into tracked classrooms, as they remain largely divided along ethnic lines (see, for example, Ansalone, 2003; Smyth & McCoy, 2001).

Over the years, myriad studies examined why the practice of tracking nonetheless continued and what relationship it may or may not have with student achievement. A number of researchers found no appreciable connection between tracking and student achievement (see, for example, Betts & Shkolnik, 2000; Kulik & Kulik, 1982). Other studies showed ability grouping had positive associations for high-level students (see, for example, Adams-Byers, Whitsell, & Moon, 2004; Neihart, 2007), while still others revealed deleterious connections between tracking and low-level students (see, for example, Ansalone, 2003; Lleras & Rangel, 2009; Oakes, 2005). The lack of consensus in previous research and continued discrepant outcomes both point to the
need for additional exploration of the topic. Furthermore, the passage of the No Child Left Behind Act of 2001 (No Child Left Behind [NCLB], 2002) and its subsequent testing and subgroup performance reporting led to a rise in tracking students by ability for instruction (Loveless, 2013).

If such a rise in tracking corresponds with changes in student achievement, policies that rate a teacher’s effectiveness by including their students’ achievement could be problematic. While some have long advocated for such accountability policies, the Race to the Top Fund (RTTT) brought the changes to fruition, as many states now include student performance data in teacher evaluations (McGuinn, 2012). Therefore, it is essential to re-examine the relationship between tracking and student achievement. Moreover, with both the increasing diversity of the American population and greater demands from the workplace, it has become more important than ever to educate all students to high levels to facilitate greater equity and opportunity for all.

As Day and Newburger (2002) demonstrated, a student’s school performance and college attendance play direct roles in his/her subsequent professions and thus socioeconomic status. Similarly, Ansalone (2003) asserted tracking’s tendency to occur along racial and socioeconomic divides meant public education essentially reproduced inputs to outputs at the same relative position within society. Effectively, a student’s socioeconomic status when he/she entered formal schooling was often his/her lifelong socioeconomic status. If tracking is associated with the exacerbation of already existing gaps in achievement, its use to further the democratic ideals espoused by American public schools comes into question.

**Problem Statement**

The primary area under investigation in this study is student-to-classroom assignment practices, namely whether students were tracked or mixed by ability into classrooms for
instruction. In this paper, tracking will be used to reference the separation of students into classrooms by subject area based on their real or perceived ability, not necessarily the assignment of students to entirely different courses of study. For example, students can take either English Language Arts (ELA) 7 or Advanced ELA 7. If those two classes were taught in separate classrooms, the students in both courses would be considered tracked, even if only for one subject. Such separation of students by ability for specific subjects has been referred to as neotracking (Mickelson & Everett, 2008). Alternatively, if both ELA 7 and Advanced ELA 7 students were taught in the same classroom at the same time with the same teacher, they would not be considered tracked but rather mixed or non-tracked for instruction. In such mixed-level classrooms, teachers work to differentiate instruction via both content and activities while maintaining a common instructional core of experiences for all students.

Given tracking’s persistence along with the controversial aspects therein, a great deal of literature already exists about tracking (see, for example, Ansalone, 2003; Betts & Shkolnik, 2000; Oakes, 2005; Slavin, 1990). Yet researchers have not reached consensus regarding the relationship between tracking and student achievement, leaving a void in the literature-to-practice arena. As such, there is a need for additional study, particularly in that studies of tracking and student achievement, as of late, are under-investigated. Aside from this void, there are two primary reasons such investigation is necessary.

First, since the institution of NCLB, student achievement test data have proliferated in ways never before seen. At the same time, statistical software continues to develop, offering more advanced and sophisticated ways in which to analyze the inherently complex data originating in schools. Tools to effectively address such data, once reserved for statisticians and econometricians, are now accessible to a wider array of researchers. Myriad variables contribute
to the variation in student outcomes and have historically conflated relationship assertions and study conclusions. Methodological advancements now allow for such variables to be statistically teased out in order to isolate variables of interest. Consequently, determining whether or not students are learning as they should be learning given various conditions, including whether or not they were tracked, is now more plausible. The intersection of data abundance and software advancement opens the door for additional educational studies that may better address and resolve the lack of consensus around tracking’s relationship with student achievement, at least from the quantitative perspective.

One relatively recent example of capitalizing on the advancement of statistical modeling techniques is the use of Value-Added Models (VAMs) to ascertain the effect a school and/or teacher has on student learning. Essentially, VAMs use vast amounts of testing data to predict or project expected student outcomes on standardized tests. If a teacher’s students collectively perform around where they would be expected to perform based on the model’s predictions, the teacher is said to have done his/her job as would be expected. Their students experienced one year of growth or learning for one year of instruction. If a teacher’s students’ collective results were statistically significantly higher or lower than expected by the VAM, they would be considered to be more or less effective than expected.

While VAM and its use for assessing the effectiveness of teachers and schools is controversial for myriad reasons, that is not the focus of this study. VAMs are nonetheless touted as the best way to quantify a student’s learning. As long as they are held up as such, it would seem VAMs could be similarly deployed to examine possible relationships and effects of such practices as tracking students by ability for instruction. While research has been inconclusive regarding the direct relationship between tracking and student achievement, as traditionally
measured by standardized test results, perhaps tracking has a definitive relationship with measures of relative student achievement, as determined via advanced statistical analyses, including VAM results. In this study, relative student achievement refers to an individual student’s performance on a standardized test with respect to his/her past performance in the same subject area. For instance, a student could perform as well as, worse than, or better than he/she had performed previously.

If, as organizationally perceived, tracking assists in the targeted instruction of students to better address their needs, sophisticated statistical analyses would yield results that, in the aggregate, would be positive for students regardless of their track. Such a finding could support the practice of tracking and allay concerns of disparate effects. On the other hand, if differential relationships are found using these advanced statistical models, including VAMs, purportedly a more discriminate metric of student progress, efforts may be advanced to eliminate the practice of tracking altogether. While tracking and its relationship with student achievement has been a focus of research for nearly a century, re-examining the relationship between tracking and student outcomes with the addition of VAMs and other statistical analyses will contribute to the assessment of classroom grouping strategies.

Second, in this era of accountability, teachers are beginning to be assessed specifically by their students’ achievement results, sometimes including VAM metrics. As high stakes decisions regarding teachers and their effectiveness are made, if the manner in which students are organized into classrooms for instruction has a relationship with students’ achievement results, that relationship must be known and considered. Under these new student achievement-based evaluations, teachers face possible consequences ranging from preferential or detrimental treatment to merit pay to firing. Given the potential consequences of erroneously attributing
some of the achievement to a teacher, rigorous examination of grouping practices and their relationship with student achievement metrics is critical.

If tracking were associated with greater relative student achievement across all levels of students, implications would arise for schools that are not tracking to begin to implement the practice. The organizational arrangement, then, would plausibly be contributing to different aggregate student achievement results for teachers assigned to mixed level classes, taking some of the ownership of the results from teachers. On the other hand, if tracking were related to lower relative student achievement results across all levels, a different set of implications would arise, though, again, the metrics seemingly would not be solely attributable to the teacher but rather, at least in part, organizationally-oriented. Yet another possible scenario would be pattern-based differential relationships between tracking and relative student achievement. For instance, tracking could be associated with better results for one level of student but worse results for another. In this case, ethical considerations would come into play as the common good of all students would need to be balanced by what may be better or worse for particular subpopulations of students. A final possible outcome would be no discernable relationship between tracking and resultant metrics, relative or not, on standardized tests. In this case, how students were organized into classrooms for instruction would be found to be independent of student performance on standardized tests. With this finding, the concerns of whether or not the organizational structure of class assignment confounded teachers’ accountability results or served to perpetuate the achievement gap could be diminished.

**Delimitations and Context of the Study**

While there were numerous variables to consider on both the input and output sides of the teaching and learning process, this study only investigated the relationship between how
secondary school students were assigned to classrooms and students’ relative achievement on state standardized tests. To assess the relationship between student-to-classroom assignment practices and students’ relative achievement, this study took place in a medium- to high-performing urban secondary school setting. Because much of the apprehension around tracking derived from apparent socioeconomic and racial segregation, the selected urban setting was purposeful. The school reflected diversity in its student body with respect to race, socioeconomic status, and academic achievement. Such diversity was important to attempt to address various concerns regarding tracking and its apparent proxy status for other variables. The choice of studying within a secondary school context was also deliberate and stemmed from the added complexity inherent to adolescent students and their dynamic interaction with teachers, affecting the classroom environment as a whole (Gamoran et al., 1995). Furthermore, the achievement gap widens in secondary schools—it would be instructive to know if a relative student achievement gap also widened and how any such movement may or may not relate to organizational structures. Lastly, tracking, as defined in this study, occurred with less frequency in elementary schools. In the primary grades, ability grouping was more typically utilized, wherein students would be grouped by ability within the same classroom as opposed to across classrooms.

Ideally a single school would have the same teachers teaching some tracked and some untracked classes. In this way, the variables of teacher and school could both be held “constant” when comparing relative student achievement results across types of grouping structures. For this study, in one grade level, each teacher taught four mixed level ELA classes and one tracked ELA class. However, the number of sections was not substantial enough to warrant the direct comparisons between grouping practices while both teacher and school were held constant. Alternatively, if a school were to employ tracking and mixing at different grades, the variables of
school and student could be accounted for in that each student would essentially serve as his or
her own control, experiencing both types of classroom assignment practices. In this case, the
selected school primarily fell into the latter category, allowing for aggregate relative student
achievement metrics to be compared across and within tracked and non-tracked classes, while
both student and school were controlled variables.

**Purpose of the Study**

The purpose of this research was to investigate the relationship between student-to-
classroom assignment practices in the secondary school setting for ELA classes and relative
student achievement on standardized state tests.

**Major Research Hypothesis**

How students are assigned to classes in the secondary school setting has a relationship
with relative student achievement on standardized state tests.

**Additional Research Hypotheses**

**Sub-Hypothesis 1:** Non-advanced students have better relative achievement when they
are assigned to mixed level classes than when they are in tracked classes.

**Sub-Hypothesis 2:** Advanced students’ relative achievement is independent of the type
of class to which they are assigned.

**Sub-Hypothesis 3:** Mixed level classrooms better reflect the school’s racial distribution
than tracked classrooms.

**Rationale for the Study**

This study considered possible consequences of school level decisions regarding how to
group students for instruction, including the distributions of both achievement and race across
classrooms. The researcher investigated the relationship between whether or not middle grades
students were tracked for ELA instruction and the subsequent relative student achievement on state standardized tests. Similarly, classroom level racial compositions were examined for differences coinciding with classroom assignment methods. It is critical for school leaders to know if the way students were organized into classes played a role in student learning. To examine this topic, quantifiable student achievement measures were utilized. In particular, state standardized tests for a cohort of students in consecutive middle grades years were analyzed against the type of grouping structure from which they resulted: tracked-to-tracked or tracked-to-un-tracked/mixed. To explore the relationship between instructional grouping practices and relative student achievement with the desired result of possible policy change, a quantitative study was warranted. As Johnson and Onwuegbuzie (2004) asserted “research methods should follow research questions in a way that offers the best chance to obtain useful answers” (p. 17).

To both disrupt the reproduction of socioeconomic class and to increase economic mobility, educational outcomes must change, pointing to the need for process changes as well. Historically, differentiation of curricula and course offerings, like tracking, began at a time when it was not necessary or expected to educate all students to high levels. Such is not the case today. If student assignment to different tracks, even in the form of different levels of the same courses, is associated with discrepant outcomes favoring the hegemonic meritocracy of old, the practice of tracking must be abandoned. At a minimum, more thorough and current examinations of student to classroom assignment practices are needed to inspect the impact of tracking students by ability not only on ultimate achievement but also on the growth of students at each level as measured by relative student achievement. Two primary benefits of utilizing relative student achievement, or growth, as a basis for educational attainment are one, students more or less serve
as their own controls and, two, growth metrics acknowledge all students do not start the year in the same positions, so both relative and temporal-based progress matter.

**Significance of the Research**

If the exploration of any educational practice, such as tracking, results in a change in policy or practice, even on a localized level, or if the unveiling of marginalization inspires further study or contributes to the understanding of schooling inequalities, the work is warranted (Mehan, 1992). Significant outcomes of this study could be changes in policies related to class placement decisions and/or instructional grouping practices altogether. As English and Steffy (2011) asserted, “educational practices that were previously found acceptable, such as tracking…, are called into question as never before” (p. 290), as student group achievement results can no longer be hidden by overall results. Yet until the light is explicitly shone on any practice and its associated possible residual effects, substantive change is unlikely. Examining tracking through the lens of relative student achievement with an eye toward VAM metrics offered a fresh and relevant perspective. If particular class assignments are related to better or worse relative student achievement results yet such metrics continue to be utilized to evaluate teachers, administrators could essentially position teachers for differential outcomes. For example, teachers who may be favored by their school administration could receive more advantageous teaching assignments, while those who may be less in favor could be assigned classes more likely to have negative VAM outcomes.

Furthermore, teachers, should they be made privy to such information, would be reluctant to take on certain teaching assignments in a school, county, or state, knowing particular situations would be less favorable from an accountability-based evaluation standpoint. Certainly, neither of the aforementioned effects would be desirable and both hinge upon classroom
assignment practices and relative student achievement metrics being dependent on one another. On the other hand, if classroom assignment practices were shown to be inconsequential to relative student achievement, or VAM results, tracking naysayers and social justice advocates would likely continue to pursue the abolishment of tracking from other angles and would likely continue to be met with inertia.

Regardless, to transform American public schools, the hegemonic practices within schools, of which tracking is but one, must be deconstructed. No longer can society afford to perpetuate class division and social reproduction by way of its institutions. Until public schools produce or contribute to more equitable student outcomes by raising the bottom- and middle-performers, every practice should be continuously evaluated for its contribution to either equity or inequity. Furthermore, as technology and other advancements become available and new tools develop to assess a practice, new investigations must ensue. Such is the case with the advent of advanced statistical analyses. With new more accessible tools, the practice of tracking must again be held up to scrutiny.

**Assumptions and Limitations of the Study**

Prior to engaging in or reading about educational research, it is important to address the underlying assumptions of the study as well as its foreseen limitations. In this study, student performance on standardized state assessments was assumed to be a legitimate way to capture the extent to which a student had learned that year’s content standards. Moreover, metrics from sophisticated analytical tools were taken as valid ways to ascertain a student’s single year relative learning gain. With respect to accountability policy implications, as long as student achievement results are utilized as part of a teacher’s evaluation, the meeting of this assumption is inconsequential. For instance, if sophisticated statistical models, including VAMs, yield
metrics inherently flawed internally yet they continue to be used as ways to hold teachers accountable, if this study finds a relationship between student-to-classroom assignment practices and relative student achievement metrics, that relationship exists nonetheless.

One of the aims of quantitative research is to utilize a large enough sample such that findings may be transferable to other contexts. The dynamic, organic nature of the teaching learning process results from myriad variables of which student-to-classroom assignment practice is just one. As such, the findings of this study may be deemed by the reader to only be relevant to the study’s immediate population at a minimum, or to similarly situated secondary schools at a maximum. This possible restriction of applicability and relevance is a limitation of this research. The focus upon a single quantitative variable of interest from three, and on occasion two, points in time further limits the study. As such, should a relationship between student-to-classroom assignment practices and relative student achievement on state standardized tests be detected, an understanding of why such an impact existed would be void. Such an understanding could emerge by expanding the study to include additional qualitative and quantitative variables in an effort to unpack the “why” behind the findings; such is not the immediate aim of this study.

Definition of Terms

The following terms will be used throughout this study. Their definitions are offered for clarity and are listed alphabetically.

**Ability Grouping:** In this study, ability grouping was defined as a within classroom method of assigning students to groups by their academic ability levels. This practice results in groups of similar ability students, though varying abilities would be present in the classroom.
Achievement Gap: Achievement gap refers to a disparity in performance between particular subgroups of students. Often the term “achievement gap” is used to talk about differences in academic achievement between White and Minority students. However, it can be used to discuss discrepancies between any subgroups with respect to any variable. When this term is used in this study, the particular meaning/definition will be specified.

Advanced Students: For the purposes of this study, advanced students were defined as those students enrolled in the advanced or honors sections of ELA. Such assignments typically reflected a student’s prior performance both in the classroom and on standardized tests.

Detracking: Detracking refers to the process of transforming a school’s organizational practices from one that had utilized tracking as its primary method to determine student classroom assignment to one that assigns students to classes without respect to student ability.

Differential Effects: Differential effects refers to the same independent variable resulting in different responses for different groups. For example, grouping students a particular way may be favorable for one group of students while simultaneously being detrimental for another group.

Education Value-Added Assessment System (EVAAS): EVAAS is a sophisticated data analysis program created by the Cary-based SAS Institute. It predicts or projects student performance on state standardized tests, thereby allowing for result-versus-prediction/projection comparisons and overall declarations of relative student learning and teacher effectiveness. It is one type of Value-Added Model and is the basis for North Carolina’s student, teacher, course, and school growth determinations.

Heterogeneous Grouping: Heterogeneous grouping is a way of assigning students to classrooms
such that a mixture of ability levels of students is represented in each classroom. For instance, this grouping scheme would result in both “honors” or advanced and “standard” or non-advanced level students learning side by side with the same teacher in the same classroom. In this study, the terms heterogeneous grouping, mixing, untracked or mixed level classes were used interchangeably.

**Homogeneous Grouping:** Homogeneous grouping is a way of assigning students to classrooms such that each classroom is comprised of one level of student—in this study, all advanced or all non-advanced students. This practice results in classrooms full of similar ability students, resulting in separate “honors” or advanced and “standard” or non-advanced level classes. The classes may or may not be taught by the same teacher, but the instruction would take place in distinct times and/or spaces. In this study, the terms homogeneous grouping and tracking were used interchangeably.

**Mixed:** In this study, mixed refers to both advanced and non-advanced students being in the same classroom at the same time with the same teacher for instruction in the same subject. Mixed is used interchangeably with un-tracked and heterogeneously grouped to refer to the same phenomenon.

**Multilevel Model:** Multilevel models offer a way to statistically model relationships among multiple variables deriving from hierarchically organized situations, such as students within classrooms. Unlike regression, multilevel models are able to address dependencies in outcomes related to higher level hierarchical features.

**Non-Advanced Students:** For the purposes of this study, non-advanced students were defined
as those students assigned to the regular level of the ELA sections. Such assignments typically reflected a student’s prior performance both in the classroom and on standardized tests.

**Normal Curve Equivalent (NCE):** Normal Curve Equivalent is a way of standardizing test scores to an equally scaled value between zero and 100. NCEs differ from percentiles in that percentiles indicate relative frequency distribution information whereas NCEs reference position on an equal-interval scale. Unlike percentiles, mathematical operations can validly be applied to NCEs. See the Appendix for a percentile versus NCE comparison. NCEs allow valid comparisons over time regardless of test difficulty.

**Peer Effects:** The concept of peer effects refers to the phenomenon of a student’s academic achievement or behavior being somewhat dependent on the academic achievement or behavior of the peers with whom a student takes classes. Most often “peer effects” is referred to in the positive, meaning students tend to have better achievement and behavior when their peers are higher achieving and well behaved.

**Relative Student Achievement:** For the purposes of this study, relative student achievement refers to how a student performed on a state standardized test one year compared to their prior year or years’ performance. Specifically, a student’s sixth/seventh grade average NCE test performance was subtracted from his or her eighth grade NCE performance. Therefore, a positive value indicated an increase in score, zero represented the same achievement as previous years, and a negative difference indicated a lower score than what was obtained the year or years prior.

**Tracking:** While there are various definitions and interpretations of tracking, in this study
tracking refers to the practice of assigning students to ELA classrooms based on their ability and/or academic achievement. Tracking may be employed in particular subject areas, such as ELA in this study, or as an overall course of study. In this study, the terms tracking and homogeneous grouping were used interchangeably.

*Value-Added Models (VAMs):* Value-added models are sophisticated statistical representations of student achievement data used to estimate how students would be expected to perform on an assessment given a typical or average educational experience. Essentially, VAMs seek to answer the question of how much value a teacher or school adds, be that positive, negative, or neutral, to students’ one year learning experience.

**Chapter Summary**

This study is both timely and relevant with regard to two main points. First, American society continues to be plagued by the achievement gap between African American and Latino students and White students. This achievement gap translates into a resource gap as students enter the workforce with discrepant educational foundations. Coupled with the continued diversification of America, it is imperative practices known to narrow or widen the gap be further investigated. By more clearly knowing the relationships between practices and outcomes, equity-based best practices can be implemented. Second, teachers in North Carolina and other states are beginning to be evaluated, at least in part, by their students’ achievement on standardized tests. If how students are assigned to classrooms has a relationship with student achievement, some teachers begin the school year already ahead or behind their peers without ever having met their incoming students. Additionally, it is important for leaders to know how such organizational decisions may relate to their school’s bottom line, namely the levels of student learning occurring from year to year. The dogma ringing through school hallways and
workrooms that it is harder or easier to obtain growth with particular “types” of students must be examined through the lens of the new accountability policies and the statistical models used therein so all stakeholders are both informed and responsible.

In the next chapter, literature relevant to this study will be reviewed. A primary focus will be on literature around the relationship of tracking with various output variables and a secondary review of literature around accountability policies.
CHAPTER TWO: LITERATURE REVIEW

In this chapter, the introduction will present a brief history of the origins of tracking, the development of achievement-based accountability for both teachers and schools, and a discussion of general educational goals and context. Four themes framed the review of the tracking literature: the relationship between tracking and student achievement, other factors related to grouping practices, grouping practices and the achievement gap, and controversies related to grouping practices in schools. Relevant accountability literature will be presented according to two themes: student achievement and teacher and school accountability and student achievement via growth and value added models (VAMs). Within each theme, literature will be presented chronologically. Finally, gaps in the literature and an overall critique will be offered.

Introduction

Since the publication of A Nation at Risk (National Commission on Excellence in Education, 1983), and the convening of the Charlottesville Education Summit in September of 1989, education has been a hot topic in both the news and political platforms. Aside from an occasional “feel-good” story about a larger than life educator making a substantive difference or the success of an irreplicable charter school, the majority of the discussion around education is that it is a problem that needs to be fixed, a failing system. The achievement gap between White and African-American and Latino students persists and in broad ways mirrors a resource gap (Rothstein, 2004). Schools, even within the same county, are markedly disparate in performance and resources, both human and material (see, for example, Darling-Hammond, 2000; Dixson &
Rousseau, 2005). Socioeconomic bifurcation persists with racial minorities disproportionately represented in the lower strata. Both experts and laypeople pontificate about the root causes and possible solutions to such disparities. One frequently cited idea is the need for young people to have positive role models in their lives, people to whom they can look for guidance in both observational and direct ways (see, for example, National Middle School Association, 2003). Most often people envision these role models to be adults, but what about their peers?

**The Problem**

It is this possible peer effect, which may contribute to disparities in outcomes related to classroom assignment practices, along with the bifurcated reality of both students’ achievement and adults’ finances that demand the exploration of the possible relationship between grouping practices in schools and student outcomes. Should a relationship be detected, it may not be solely the result of the grouping practice but rather a sequence of factors that relate to whether students are low or high tracked or mixed together, such as teacher expectations, classroom culture, teacher quality, etc. (Oakes, 1985; Watanabe, 2008). Nonetheless, all of the aforementioned components, along with others, have been associated with student assignment practices. Consequently, in this study, how students were assigned to classes served as the analytical umbrella. What are the consequences of whether or not students are heterogeneously or homogeneously assigned to classes according to ability? Given both policy shifts and persistent societal stratification, understanding the relationship between tracking and relative student achievement is critical to both accountability measures and to social justice. Inasmuch as the research has shown tracking may benefit top students, it has also been shown to be detrimental to the lowest students, perpetuating society’s social constructs (Ansalone, 2003; Conger, 2005; Oakes & Guiton, 1995). With the advent of policies requiring the use of student achievement
data to rate teacher and school effectiveness, what is the relationship between how students are
assigned to classes and relative student achievement?

On a continual basis, various stakeholders within the school community make decisions
regarding how to organize students for instruction. For educational leaders, decisions regarding
student assignment practices are influenced by ideological beliefs about student learning,
parental pressures, and prior experiences. While not specifically examined in this study, a type of
tracking or ability grouping often occurs within classrooms. Teachers, independent of an
administrators’ setup, frequently choose to group students by similar or differing ability. Parents
and students, during the course selection process and at times dependent upon the class
organization structure within the school, select levels of a given course. What impact, if any, do
these decisions have upon the common goal of all parties involved, student learning?

Ideally, the research would synthesize to an area of common ground from which all
stakeholders could operate. However, conditioned by both beliefs and context, what may work in
theory often differs in practice (Nuthall, 2004). Consequently, the primary goal of reviewing the
literature on student to classroom assignment practices is a grounded understanding of the beliefs
and issues underlying opposing views. It is essential for administrators, teachers, parents, and
students to understand the implications their choices regarding class placement have on both the
greater population and on the individual student. Please note, in an effort to maintain consistency
throughout this review, regardless of how authors chose to refer to student to classroom
assignment practices, henceforth tracking will indicate separating students by ability across
classes while ability grouping will indicate grouping students by ability within classrooms.
Furthermore, this study specifically investigated tracking versus mixing across classes.
The Historical Background of Grouping Practices

The origination of tracking as a school response to increased diversity likely conditioned present day practices and beliefs. Initially, tracking practices addressed immigration. Later, tracking was justified as a way to better prepare students for different post high school options. While today there is a push for all students to be “college and career ready” at graduation, tracking nonetheless persists.

Kulik and Kulik (1982) connected today’s tracking practices to the Santa Barbara Concentric Plan first implemented at the start of the nineteenth century. In that plan, grades were separated into three different sections, all of them learning the same fundamental curriculum, but two sections learned incrementally more than the baseline (Kulik & Kulik, 1982). Such stratification abounds in today’s school organizational structures and policies. At the beginning of the twentieth century, schools tracked students in an effort to create a more tailored workforce. Over time, tracking became less popular until the landmark Brown v. Board of Education (1954) case, after which tracking came back into prominence. Resistant to and irrespective of desegregation orders, schools utilized tracking to maintain segregation (Chayt, 2010). Throughout the 1990s, concerns over equity drove many detracking efforts. However, following NCLB and performance-based accountability measures, schools began separating students by ability once more, resulting in disproportionately White higher level classes and disproportionately minority lower level classes (Jackson, 2009; Kozol, 2006; Lleras & Rangel, 2009; Loveless, 2013; Oakes, 2005).

Educational Goals and Context

Two common goals within public education are to obtain high levels of student achievement and to decrease achievement gaps, which together support a more equitable society
(Haycock, 1999; Wagner, 2008). Over the last five to ten years, measures of student achievement have shifted to include a gauge of how much a student learns or grows in one year, which acknowledges the varied starting points of students with respect to ability and prior knowledge. This type of growth measure can be obtained via the use of sophisticated statistical models, including Value Added Models (VAMs). Because tracking occurs in most public schools today, it is important to understand its relationship with both the achievement gap and student achievement, including the newer metrics of student learning (Ansalone, 2010; Archbald, Glutting, & Qian, 2009; Kozol, 2006; Loveless, 2013; Mickelson & Everett, 2008; Oakes, 2008).

Often tracking begins in response to students’ perceived level of readiness. As such, what students already know coming into a grade or class is of primary importance. It influences not only which classes they take but also at what levels. Students’ prior knowledge and performance also determine whether they receive extra support. Often such remediation takes the form of an additional class period, thereby taking away an otherwise elective class for low performing students.

Ultimately, how students perform in school has a direct impact on whether or not they attend college and their choice of livelihood, thereby playing a key role in their socioeconomic status (Day & Newburger, 2002; Jenks, 1998). If American public schools truly served as the great equalizer and provider of equal opportunity to all, they would act to close achievement gaps due to factors related to parenting and access that exist for students upon entry to public schools. Assuming its most altruistic intention, tracking seeks to serve students in the most equitable way possible by meeting students at their level. It is of utmost value for all stakeholders to know and understand the conceivable impact of this widely used practice and to assess whether or not it accomplishes its goals. By examining the beliefs and attitudes behind
tracking along with research about its relationship with student achievement, guidelines for practice can develop.

The major focus of this literature review will be on the effects of tracking and similar ability grouping practices from the perspective of how students are assigned to classrooms. Tracking is the primary focus due to the extent of the literature and that student to classroom assignment practice is the principal independent variable of interest in this study. Four themes provided a framework to examine the literature: the effects of grouping practices on student achievement, grouping practices and the achievement gap, controversies related to grouping practices in schools, and other factors related to grouping practices. A smaller focus of this review will be around literature related to shifts in educational accountability policies. These policy shifts will be a smaller component of the review, because the primary interest of the study was to examine the relationship between classroom assignment practices and relative student achievement. The accountability policy shifts make the study more salient, with possible implications not only for educational equity but also for the validity of new teacher evaluation policies.

While myriad studies have been done on some element of the topic of tracking and student achievement, the review will highlight the need for additional study around the relationships between arranging students for instruction via tracking or mixing and student achievement as measured by sophisticated statistical metrics. In the last few years and largely stemming from Race to the Top grant acquisition, growth-based student achievement metrics are being utilized to assess teacher and administrator effectiveness. As such, it is important to more fully investigate the relationship between student classroom assignment practices and relative student achievement metrics.
Themes Through the Literature

The Relationship Between Tracking and Student Achievement

**Overviews of research.** A number of researchers have reviewed vast amounts of literature related to tracking and its relationship with student achievement. Several are highlighted here to provide a baseline for the many additional studies and analyses that have taken place since these examinations occurred.

In their review of research, Kulik and Kulik (1982) specifically examined only experimental studies that tracked students into classes by ability. Additionally, whereas many reviews relied extensively on narrative results and simplistic statistical “boxscore” analyses, Kulik and Kulik (1982) employed a meta-analysis in order to provide objective quantitative results. Of the 52 studies reviewed, Kulik and Kulik (1982) reported 36 of them indicated tracked students outperformed non-tracked students, yet only eight of these 36 were statistically significant. They also reported fourteen studies showed non-tracked students achieved at greater rates than tracked students, with only two being statistically significant. One study found essentially no difference in achievement between the two groups (Kulik & Kulik, 1982).

Altogether, the effect size was 0.10, meaning tracked students achieved, on average, one tenth of a standard deviation higher than their non-tracked counterparts. Such a difference corresponded to an average performance at the 54th percentile versus the 50th percentile (Kulik & Kulik, 1982). The variation around the average of 0.10 substantiated a belief that factors other than grouping structures affected achievement results. As such, Kulik and Kulik (1982) discovered the population studied played an important factor: the 14 studies involving programs specifically tailored for gifted and talented students yielded the most significant and positive effect sizes. Conversely, the four programs designed for remedial students had near zero effect sizes, as did
the 33 studies involving non-restricted populations. Nineteen of the 33 unrestricted studies included enough information to determine effect sizes by ability level, none of which were significant alone or when compared to each other (Kulik & Kulik, 1982).

In 1990, Robert Slavin also synthesized the tracking literature. At that time, not much research comparing tracked students’ achievement to heterogeneously-grouped students’ achievement had been done since the early 1970’s. Instead, studies had compared student achievement between entirely different tracks or courses of study. Slavin (1990) examined results from 29 secondary school studies, among which there were six randomized experiments, nine matched experiments, and 14 correlational studies. The reviewed studies spanned a 60-year period. Across all studies, the impact of ability grouping on student achievement was negligible, with non-statistically significant effect sizes and a median effect size of zero. Of note, when researchers compared achievement gains between tracks, high track students accelerated in achievement when compared to lower track students’ achievement. Slavin (1990) posited significant differences in initial ability and motivation among high and low tracked students were too substantial to be sufficiently accounted for in statistical models. As such, comparing tracked to mixed students’ achievement was more reliable than comparing high- to low-track students’ achievement (Slavin, 1990).

Gamoran (1992) also provided a synthesis of the tracking and ability grouping research. He questioned Slavin’s (1990) conclusions about such practices having no overall impact on student achievement, arguing that while the average of all the studies was essentially zero, there also was significant variability in the effects found within individual studies. Gamoran suggested this variability may be due to how schools were engaging in internal processes regarding student grouping decisions rather than just statistical anomalies, as handled by Slavin. Gamoran (1992)
viewed this masking of results as a theme throughout the literature, discussing both international and American studies that reported tracking’s overall non-effects on student achievement. Yet when the data were more closely examined, it became evident tracking students by ability yielded higher gains for the higher tracks and lower gains for the lower tracks when compared to similarly able students in non-tracked situations. While the overall results had appeared ineffectual due to aggregation, in fact tracking exacerbated gaps between high and low track students. Furthermore, in the high school setting, the longer students were separated in curricular tracks, the greater the inequality in subsequent achievement.

After examining data from 20,000 students in grades ten through twelve, Gamoran (1992) discovered gaps on achievement tests between seniors in different tracks, for example, vocational or academic, were wider than achievement gaps between students who dropped out of school and those who remained. While recognizing this reporting of overall results was that of which Gamoran (1992) was hesitant, the implications were nonetheless startling. Overall, then, Gamoran concluded tracking rarely impacted a school’s overall achievement level, yet that same aggregate achievement was distributed in disparate ways that increased inequalities both within and outside of school. As such, tracking should be minimally utilized and in cases where the practice was to persist, improving upon the ensuing instruction within all tracks needed to be an area of focus (Gamoran, 1992).

**Grouping Practices in Elementary schools.** Of note, while the present study focused on relationships between student to classroom assignment practices in the middle grades and relative student achievement, reviewing similarly focused research from the elementary school level can add to a baseline of knowledge regarding practices and their possible effects leading up to secondary school education.
Similarly, while this review focused on the effects of between class tracking, the following study was included for its relevance regarding student opportunities to learn, which may be affected from both a breadth and depth standpoint when students are tracked. Sorensen and Hallinan (1986) studied the effects of within classroom ability grouping on elementary school students’ reading achievement. Sorensen and Hallinan (1986) posited a different model for learning that included the interaction between students’ opportunities to learn and their ability and effort toward learning—quite simply, if students were not afforded the opportunity to learn an aspect of the curriculum, they were not going to learn it. When students were grouped by ability within classrooms, they received less direct instructional time from the teacher than if they were instructed whole class. With that in mind, Sorensen and Hallinan (1986) wondered if the loss of instructional time had an adverse effect on student learning. They found ability grouping had no impact on either individual reading achievement or reading growth over the course of a year. However, they noted under their learning model, ability grouping would only have an effect if it served as a proxy for ability and effort (Sorensen & Hallinan, 1986).

Nevertheless, in a statistically significant way, ability grouping resulted in fewer opportunities to learn, though higher groups had more opportunities than lower groups (Sorensen & Hallinan, 1986). Essentially, then, while ability grouping yielded no achievement impact, it did decrease opportunities to learn material. Interestingly, race was a significant factor on student achievement in classes employing ability grouping but not in heterogeneously grouped classes. These differences were negative for African American students in ability-grouped classes, yielding a compounded impact due to African American students’ higher frequency of being in ability-grouped classes. Importantly, this finding indicated ability grouping was not a neutral
practice, but rather one that amplified inequity with respect to educational outcomes (Sorensen & Hallinan, 1986).

Slavin (1987) utilized a best-evidence synthesis to analyze myriad studies about within- and between-class ability grouping in elementary schools. Consistent with other work, Slavin (1987) found little to no relationship between tracking students by ability and student achievement, though ability grouping within math classes had a positive effect on achievement. The greatest overall student achievement occurred in schools where students were tracked for one or two subjects and with heterogeneously grouped or mixed-level classes for the majority of the day (Slavin, 1987). Furthermore, if grouping were instituted, Slavin (1987) concluded the methods used for establishing groups should be revisited frequently and allow for flexibility of assignment.

The following study, though taking place in a developing country, was included for both its explicit experimental design and the clear results, both of which are atypical in the literature. Duflo, Dupas, and Kremer (2009) led a study in Kenya in which 60 schools randomly assigned, without regard to ability, first grade students to one of two classrooms, while another 61 schools tracked students to either a higher or lower ability class. The study’s funding accounted for a second teacher in each case, thereby substantially lowering class sizes in both conditions. As such, the achievement of all students was expected to increase more than prior classes (Duflo et al., 2009). While the results of the study could only validly be applied to schooling in developing countries, the outcomes were clear: overall, students, regardless of initial ability, benefited from the tracked class structures more so than their untracked peers (Duflo et al., 2009). Interestingly, students originally in the middle of the ability spectrum, with some assigned to the ‘high’ class while others went into the ‘low’ class, fared similarly to one another (Duflo et al., 2009). While
gains were similar in both the high and low classes and consistently higher than the non-tracked classes, the lower classes had most of their gains in lower level concepts while the higher classes gained mainly in the higher level concept attainment (Duflo et al., 2009).

Lleras and Rangel (2009) examined the reading achievement for students in grades one through three to ascertain the impact of within class ability grouping, specifically as it related to the reading performance of African American and Hispanic students. They found African American and Hispanic students placed in lower ability groups had significantly lower gains compared to non-ability grouped students. Conversely, students placed in higher ability groups had substantially higher gains than those not grouped at all by ability. Putting these two results together, Lleras and Rangel (2009) resolved gains by the higher-level students did not offset the losses by the lower level students and posited such early-years gap-widening practices may have significant trajectory implications. Their conclusion again illustrated the need to examine overall apparently neutral results for potentially differential results by student groups.

**Grouping Practices in Secondary schools.** Of particular interest to this study was the literature on how tracking related to student achievement in secondary schools. While the overviews of research previously presented (Gamoran, 1992; Kulik & Kulik, 1982; Slavin, 1990) also reviewed tracking’s impact on secondary school students’ achievement, the following specific studies are presented in more detail to add context.

By examining the math achievement and high school graduation rates of nearly 11,000 students across the United States via the High School and Beyond Survey, Gamoran and Mare (1989) noted achievement gaps were greatest when students were separated by tracks for the full instructional day. This finding held even when initial differences were controlled. The overall effects of tracking, then, were not neutral and contributed to inequality. Both achievement gaps
and high school graduation rates grew wider and farther apart the longer students experienced tracking (Gamoran & Mare, 1989).

Hallinan (1994) explored possible differential achievement effects of tracking across schools, a previously unstudied element. She believed the practice would continue, so the best possible policies should be known and replicated. The longitudinal study surveyed tracking practices from the 7th to 9th grades, starting in 1986, across two cohorts totaling 4,563 students. Multivariate analyses revealed significant differences in how schools assigned students to tracks across schools with students assigned to higher tracks achieving higher levels of learning (Hallinan, 1994). Hallinan found tracking impacted student achievement in two primary ways: first, the quality and quantity of instruction increased with the track level, and, second, social and psychological impacts on students affected motivation and effort, again increasing with track level (Hallinan, 1994). The effects of tracking on student achievement varied substantially both at different schools and within the same school. Different tracks had different outcomes, always favoring the higher tracks, resulting in additional inequities (Hallinan, 1994).

By highlighting the deleterious effects tracking had on minority and lower socioeconomic students, Oakes (1985) provided some of the foundational work that brought concerns of equity to the forefront of education reform efforts. Both the social and educational climates in the early to mid-1990’s included a view that detracking could serve as a possible panacea to both public school woes and student achievement challenges. This optimistic sentiment largely stemmed from the results of Slavin’s (1990) review, which indicated tracking had a flat relationship with student achievement. If tracking provided no achievement advantage, the social benefit of detracking would not negatively affect achievement. However, Argys, Rees, and Brewer (1996) cautioned against practice-impacting conclusions drawn from Slavin’s work, citing his use of
small single school samples, unpublished dissertations, and the ages of some of the studies synthesized. In the first study to utilize statistical modeling to examine the impact of tracking on student achievement while controlling for many variables, including possible resource allocation inequities, Argys et al. (1996) found detracking was not as clear cut a solution as previously presented. Using data from the National Education Longitudinal Study of 1988, which included achievement information from 8th and 10th grade along with linkages to the student, teacher, and class, they identified mixed results with respect to tracking’s impact (Argys et al., 1996). Moving from tracked classes to heterogeneous classes, the model predicted low ability students would gain over eight percentage points in math achievement, while high ability students would have a similar but slightly smaller decrease in math achievement, and average students’ achievement would decrease by approximately two percentage points (Argys et al., 1996). This time, the predicted differential impact favored lower track, often disadvantaged students, perpetuating murky guidance for practice.

Betts and Shkolnik (2000) sought to improve the comparability basis by including ability controls even for heterogeneously grouped students in mixed-level classrooms. They took a school finance analytical approach, focusing on possible resource allocation implications and the relationship between inputs and outputs involved therein. Betts and Shkolnik (2000) utilized a model referred to as the “prototypical education production function,” for all cases, including additional controls for initial class ability to determine if tracking had a positive average impact on student achievement and if there were differential effects depending upon the initial ability level of the class. Betts and Shkolnik (2000) utilized data from the Longitudinal Study of American Youth, which included national cohort data following groups of students from grades seven to nine and ten through twelve from 1987 to 1992.
Overall, Betts and Shkolnik (2000) found no appreciable differences in achievement between schools employing tracking versus those utilizing mixed-level classes for instruction. When they examined results for possible differential effects of tracking, they compared students in non-tracked or mixed classrooms with similar ability students in tracked classrooms. They did not find any negative effects for lower level tracked students, slight negative effects for middle level tracked students, and slight positive effects for higher level tracked students. The differences of note were that prior studies found negative low effects and greater positive high effects, possibly due to less sophisticated controls for initial ability. Overall, then, Betts and Shkolnik (2000) concluded formal grouping policies, when controlled for initial ability, had no overall effect on student achievement, differential effects existed but in much smaller amounts than previously presented, and the practice of tracking did not appear to have an inequitable impact on resource allocation.

Similar to Sorensen and Hallinan’s (1986) discussion of the impact of ability grouping on students’ opportunities to learn, Ansalone (2001) referenced significant gaps in curricular coverage, dependent upon assigned track. Teachers of higher tracks managed to cover approximately 85% of the curriculum while lower track teachers of the same class only taught about 60% of the curriculum. Clearly, then, lower track students missed significant amounts of material altogether, further disadvantaging them and hindering them from advancing to higher levels in subsequent years. In that lower track students’ opportunities to learn were restricted, tracking inevitably perpetuated social structures (Ansalone, 2001). Ansalone (2003) also reviewed research related to the practice of tracking from American and British sources. He considered a number of areas, such as whether or not social class and racial bias influenced tracking, whether or not tracking affected students’ academic achievement or self-esteem, and
whether a tracking by-product could be different curricula altogether. By culling both countries’
research together, he found tracking could be particularly detrimental to lower level students who
did not have strong family foundations related to support and value of education and who tended
to lack access and skills from the beginning (Ansalone, 2003).

**International studies related to grouping practices.** Schools in Europe utilized three
types of grouping practices: streaming, wherein students were separated by achievement score
into the same track level for all classes; setting, in which students were separated by ability
differentially, allowing for high level classes in one subject but lower in another for the same
student; and mixed ability grouping, which sometimes happened by chance and other times was
specifically designed such that all levels were in the same classroom at the same time (Smyth &
McCoy, 2001). For comparison purposes, what they referred to as setting would be
commensurate with tracking, as defined in this study. In tracked schools, in addition to lower
expectations and lower quality instruction, low tracked students received significantly lower
grades, leading to greater inequality. This discrepancy existed even though the overall average
scores were similar to non-tracked schools (Smyth & McCoy, 2001). In total, Smyth and McCoy
(2001) found strong evidence to lessen the extent of rigid separation of students by ability.
Resulting in similar findings, William and Bartholomew (2004) studied the math achievement of
955 students at six different London schools over the course of four years. They found tracking
students into separate classes by ability had a significant influence on student performance, with
the highest students doing the best and the lowest grouped students faring the worst (William &
Bartholomew, 2004).

Taking a broader perspective, Van de Werfhorst and Mijs (2010) wanted to know the
implications of institutional practices for national achievement and whether the tension between
educational efficiency and educational equality was warranted. Invariably, countries operated
different educational systems. For instance, some countries had comprehensive public schooling
throughout a student’s experience, while others only offered the same schooling until a certain
grade at which point either aptitude/achievement or interest determined a student’s school
placement, which essentially created tracked schools (Van de Werfhorst & Mijs, 2010).
Countries operating tracked schools had greater inequity in student achievement, more variability
among students, and higher dependence of achievement on race/ethnicity and socioeconomic
status than countries without tracked schools (Van de Werfhorst & Mijs, 2010). Tracked schools
also resulted in greater inequity in opportunities for students, and, importantly, lower overall
average student achievement (Van de Werfhorst & Mijs, 2010). In contrast, countries supporting
a more standardized educational experience for their youth had more equitable opportunities for
students and higher average student achievement (Van de Werfhorst & Mijs, 2010). In this case,
equity and efficiency were not in opposition as previously believed. If this finding were to hold,
important policy implications could ensue, should the window open: instead of battling
ideologies of equity against those of efficiency, both could be advanced simultaneously with
energies directed more productively at improvement.

Other Factors Related to Grouping Practices

While this section of the literature review will center on the discussion of variables not
immediately relevant to the current study, its inclusion was purposeful. Each study offered
insight into the foundations for some of the controversy around the practice of tracking and/or
related to issues previously shown related to student achievement. The former was important
should an increase in detracking efforts be indicated from this study and the latter informed a
possible step between the assignment of students to classes and their learning attainment.
Grouping practices and students’ self-concept. In addition to examining the relationship between tracking and student achievement, a number of studies have focused on or included how tracking affected students’ self-concept. Many people believed self-concept impacted students’ efficacy and therefore engagement and achievement. As previously described, Kulik and Kulik (1982) investigated tracking versus student achievement. Of the studies they examined, fifteen also included information regarding students’ self-concepts, yielding mixed and mostly non-significant results. The average effect size was just 0.01.

Through interviews, observations and specialized assessments, Boaler, William, and Brown (2000) found tracking students by ability for mathematics instruction had negative socio-emotional effects for students in all levels of grouping. Students in higher tracks often felt increased pressures, expectations, and a pace incommensurate with their understanding, all detrimental to their confidence and achievement in mathematics (Boaler, William, & Brown, 2000). Similarly, though on the opposite side of the spectrum, students in lower tracks described being bored, not challenged, and feeling unimportant based on low expectations and teacher practices (Boaler, William, & Brown, 2000).

Ireson and Hallam (2009) surveyed more than 1600 students from eight mixed-grouped, eight combination-grouped, and seven tracked-grouped schools to examine possible effects of grouping structures on students’ academic self-concepts. Their study took place in England with a baseline dataset collection followed two years later by a re-collection from the same students (Ireson & Hallam, 2009). They found an inverse relationship between the extent to which schools utilized tracking and students’ academic self-concept, defined as their perceived competence, interest, and enjoyment of a subject or school (Ireson & Hallam, 2009). In other words, the heterogeneously grouped/untracked schools’ students had the highest average
academic self-concepts while the purely tracked schools’ students had the lowest. Within tracked schools, though the effect sizes were small, statistically significant differences occurred between groups with the highest academic self-concepts in high ability groups and the lowest academic self-concepts in the lowest groups—this was true in each of English, math, and science (Ireson & Hallam, 2009). In performing various rounds of ANOVA and controlling for myriad variables, Ireson and Hallam (2009) found the socioeconomic status of students was not related to their academic self-concept. It was important to understand the relationships between grouping structures and students’ academic self-concepts because academic self-concepts impact students’ intentions for learning, which in turn affects their achievement (Ireson & Hallam, 2009).

**Grouping practices and peer effects.** Public policy debates about ability grouping and desegregation also included discussion of peer group effects (Hanushek, Kain, Markman, & Rivkin, 2003). There appeared to be an assumption that around whom a child was learning would affect their learning, yet few studies delved directly into an explicit examination of these supposed peer effects (Hanushek et al., 2003). Hanushek et al. (2003) utilized a matched panel data set wherein similar students from different groups were matched with each other to statistically control for as many potentially confounding variables as possible, seeking to determine the validity of the peer group effect. Their “fixed effect framework” controlled for systematic school and family influences. The use of lagged peer achievement values in their model, if anything, underestimated peer effects, as some students changed substantially over time (Hanushek et al., 2003). The data set came from the University of Texas at Dallas Texas Schools Project and included three cohorts, each with over 200,000 students from over 300 schools, tracking whole grade achievement from grades three through six, starting in 1992 (Hanushek et al., 2003). Hanushek et al. (2003) found positive peer effects across the entire test
score distribution, ranging from 0.15 to 0.24, and believed their results to be a lower bound due to lagged test score substitutions. These effects were strong, indicating peer achievement was a significant factor in student learning.

Viewing peer effects from a school level, Jackson (2009) found students attending better schools with higher-achieving peers benefited in several ways, such as higher test scores and additional years of subsequent education. Students from all levels benefited from being with higher-achieving peers. Girls benefited the most, though low-achieving students realized the least amount of benefit. When schools were essentially tracked by ability, the quality of the schools varied greatly, perpetuating and expanding academic achievement disparities (Jackson, 2009). Boucher, Bramoullé, Djebbari, and Fortin (2010) discussed the importance of knowing possible peer effects on student achievement, indicating such effects were difficult to discern due to challenges regarding the control of myriad potentially confounding variables. Therefore, Boucher et al. (2010) utilized an econometric model derived from Lee’s (2007) work that sought to identify interaction effects. They collected math, science, history, and French achievement test results in Quebec, Canada, from 116,534 fourth and fifth grade students (Boucher et al., 2010). Boucher et al. (2010) found significant and positive peer effects in math, French, and history, with math being the largest at 0.83, meaning for every point a student’s average peer achievement increased, an individual student’s achievement was expected to increase by about 0.83 points as well.

**Teachers’ methodology and expectations.** A number of studies indicated the actions of the teacher in a tracked versus untracked classroom contributed to the student achievement effect more so than the grouping practice itself. When Slavin (1990) found no appreciable effects of tracking on student achievement, he was surprised, likening the results to studies showing class
size did not have an effect on student achievement until ratios were prohibitively low. Slavin (1990) posited both cases might be indicators that how schools were organized may not be as important as the pedagogical practices employed therein. Both Ansalone (2003) and Gamoran (1992) agreed any effect of tracking was less derived from the explicit act of separating students and more from the subsequent instruction, which included teacher expectations, academic press or focus, and climate.

Studying students aged 8 to eleven, Ben-Ari and Kedem-Friedrich (2000) found no matter the setting, when teachers used ‘complex learning techniques,’ greater cognitive development followed. Complex learning techniques involved teachers encouraging and supporting students in verbal interactions and students’ interdependence with each other related to new and acquired content (Ben-Ari & Kedem-Friedrich, 2000). How much time spent on task with social interaction and how much teachers facilitated such activity were both positively related to cognitive development. Boaler, William, and Brown (2000) supported such an instructional focus, finding tracked classes varied less in pedagogical practices than mixed-level classes, resulting in overall negative effects on all tracked students. Similarly, Gamoran et al. (1995) posited tracking students by ability served as the entry point to what mattered most: the ensuing instruction. Regardless of the grouping structure, if student participation and engagement were optimized, off task behavior minimized, and the teacher planned so as to deliver explicitly coherent lessons, student learning could be maximized. They conceded, however, that maintaining all of these factors was more challenging in lower level tracks (Gamoran et al., 1995).

Taking a school rather than teacher expectations approach, Lee and Bryk (1988) examined factors related to the social distribution of high school math achievement. They
utilized data from 2,050 Catholic school (CS) students from 83 schools and 1,883 public school (PS) students from 94 schools. They wondered if structural differences between the two types of schools explained, at least in part, math achievement differences between CS and PS students (Lee & Bryk, 1988). CS students utilized tracking, and students were two times more likely to be in the higher demand academic track than PS students, perhaps indicating a difference in expectation (Lee & Bryk, 1988). Structurally, Lee and Bryk (1988) found CS had different and more prescribed graduation requirements for all tracks, resulting in their non-academic student course-taking patterns looking more similar to the college prep track than analogous tracks in PS. They concluded differences between CS and PS tracking and course patterns were institutional in nature (Lee & Bryk, 1988). Lee and Bryk (1988) found a stronger “institutional pull…toward academic pursuits” in CS, whether it was initial track placement not determining subsequent course taking, or academic background being less critical for initial placement, or the similar graduation requirements established for all students. The CS staffs differed from the PS staffs as well, being more likely to share a common mission, vision, and goals for all students to advance as far as possible (Lee & Bryk, 1988).

In Europe, studies showed lower streams (or tracks) led to lower academic demands and less engaging instructional techniques, culminating in students having negative views of their own abilities (Smyth & McCoy, 2001). Furthermore, in tracked schools, pre-existing student differences were reinforced rather than ameliorated, including the overrepresentation of minority and low SES students in lower tracks (Smyth & McCoy, 2001). Regardless of the level of the class, in tracked classrooms, teachers tended to utilize fewer strategies of instruction than teachers of heterogeneous, mixed ability classes (William & Bartholomew, 2004). Additionally,
William and Bartholomew (2004) found teachers of lower level classes overall had lower expectations of their students than teachers of higher level classes.

Ansalone and Ming (2006) suggested individualizing instruction as a way to ameliorate some of the differential backgrounds of students. One successful intervention to address disparate seventh grade learning needs in Bermuda was the implementation of a Programmed Learning Sequence (PLS), which utilized computers to individualize instruction. Overall, students involved in the study scored higher and showed preference to the PLS methodology (Ansalone & Ming, 2006). While using such a specific program may not be a practical consequence of the results, greater efforts to match instruction to students’ learning styles would likely result in increased achievement, thereby assisting in eliminating gaps between students. This kind of extreme tracking may become more prevalent with the advent of educational technologies. If so, it will be important to monitor for potential differential impacts over a longer period of time.

**Student and teacher assignment decisions.** If which class or level of class a student takes and which teacher a student has both affected the student’s academic achievement, how were such student to teacher assignment decisions being made? While elementary schools generally utilized within class ability grouping, secondary schools more often employed between class tracking (Loveless, 1998). Such grouping in high schools frequently resulted in entirely separate tracks, which required different levels and types of coursework to satisfy graduation requirements (Slavin, 1990). Sometimes students were assigned to blocks of classes all of one ability. In other instances, students took higher-level courses for just one subject (Mickelson & Everett, 2008). Even absent full course of study separation, the vast majority of high schools tracked students by ability into different levels of English and mathematics classes, while many
schools also tracked in other subjects (Hallinan, 1994). School level decisions affected tracking efforts both by how they created tracks and by how they assigned students to them. Schools also played a role in determining the learning opportunities associated with and provided within each track (Hallinan, 1994). The size of the school affected the number of tracks offered, which in turn affected the level of homogeneity within each track. Additionally, local priorities could play a role in track composition in that a principal may have a goal of racially heterogeneous tracks or of academically homogeneous tracks, both of which would affect students’ assignments (Hallinan, 1994; Lee & Bryck, 1988). Even within the same school, different criteria could be used for different track assignments. Due to variations in assignment policies, student populations, and staffing, two students assigned to the same track or level of class at different schools did not necessarily mean the same thing (Hallinan, 1994).

A study of three comprehensive high schools in similar locations but with varying demographics more closely examined how high schools made decisions regarding which students take which classes on which tracks (Oakes & Guiton, 1995). Oakes and Guiton (1995) found all three high schools offered both academic, college-preparatory classes along with vocational, workforce-preparatory classes but the distribution and mix of the two types varied greatly among the schools. The most affluent community’s high school offered far more academic-based and higher level courses, while the school serving the poorest community offered a much greater number of vocational courses (Oakes & Guiton, 1995). All three schools made substantial efforts to place students into the appropriate set of courses upon entry by gathering historical achievement information and teacher recommendations. However, at each school, a student’s movement away from this initial placement seemed possible but not likely or expected (Oakes & Guiton, 1995).
Overall, teachers and administrators indicated seeing race and social class as indicators of ability as they referenced changing demographics as the primary rationale behind changing course offerings. Students in higher track classes also tended to have more supportive and advocating parents who, together, benefited from a more tightly put together package of courses directly leading to post-high school opportunities (Oakes & Guiton, 1995). In contrast, lower track students were further disadvantaged by a looser curriculum that required more guidance to successfully navigate. In such tracks, it was less clear how trajectories, sequences, and combinations of courses interplayed to position a student for graduation. While students and parents were given some choice with regard to class and track selection, they, along with school personnel, tended to choose placements resulting in a disproportionate number of minorities in lower tracks (Oakes & Guiton, 1995). Such results suggested the consequences of tracking stemmed from a complex array of factors deeply rooted within society.

Argys et al. (1996) examined teacher characteristics in an attempt to explain differential outcomes. They found neither teacher experience nor degrees beyond bachelor’s were significant factors in student achievement in any setting, while having a certified math teacher was a positive factor, particularly in the below average tracks. William and Bartholomew (2004) found there was a systematic difference in the type of teachers assigned to the different levels, with lower qualified teachers being tasked with teaching lower ability classes while highly qualified teachers were assigned to teach the higher level classes.

Archbald and Keleher (2008) asserted schools needed accurate and appropriate data at their fingertips in order to be able to group students correctly, if at all. Additionally, if tracking were employed, there should be flexibility built into the organizational structure to allow for some levels of freedom for students to move among and between tracks. Archbald and Keleher
(2008) further suggested having data was the first step to enhancing school organization, but understanding the trends within the data and utilizing it consistently were critical subsequent steps in school improvement efforts. As teachers played a critical role in placement recommendations for their students in the following year, to be fair, they needed to be informing their decisions with more than students’ current year performances, augmenting their picture with historical data. Administrators could not ‘fix’ their schools by employing an outside program or research without first having a deep understanding of the inner-workings and underpinnings of their school’s data, climate, and culture.

**Grouping Practices and the Achievement Gap**

*Grouping practices and their influence on different levels of students.* Given the social inequities stemming from the achievement gap between White and minority students, it is important to know how tracking may be playing a role in closing or widening the gap. A number of studies showed minorities were disproportionately assigned to lower tracked classes, while the overall impact of tracking by ability favored higher-level students but was detrimental to lower level students, thereby exacerbating differences present at the start of the year (Argys et al., 1996; Hallinan, 1994; William & Bartholomew, 2004). Lleras and Rangel (2009) also found differential effects on the reading achievement of tracked students in grades one through three. Specifically, African American and Hispanic students placed in the lower tracks had significantly lower gains compared to similar matched non-tracked students. Conversely, students placed in the higher tracks had substantially higher gains (Lleras & Rangel, 2009). Even though Duflo et al. (2009) found between class tracking benefited all levels of students, closer examination revealed disparate gains. Higher tracked students gained in higher level concepts and lower tracked students’ gains stemmed from lower level concepts. Betts and Shkolnik
(2000), however, found when initial ability was controlled, tracking students by ability had no appreciable effects.

**Equity concerns stemming from grouping practices.** Public schools were founded upon democratic ideals to serve as the great equalizer of opportunity. Therefore, practices resulting in inequitable outcomes with respect to a particular variable, such as race, are of considerable interest. Concerned about “lower-class” students attending inferior schools, Coleman et al. (1966) investigated whether or not the quality of a school impacted the achievement of its students. They surveyed over 4,000 public schools in the United States, including 60,000 educators and 570,000 students. Coleman et al. (1966) discovered the actual school attended mattered less than a student’s outside of school factors and the attitude they brought to school. They found variations in student achievement were greater within schools than between schools, lending credence to the school itself having less to do with the outcomes than the students (Coleman et al., 1966). Alternative explanations for the internally discrepant achievement may be the processes and structures within the schools, such as tracking. Coleman (1967) followed up his team’s findings with an investigation into how to systematically ameliorate the inevitably inequitable educational opportunities of students from diverse backgrounds. Distributing resources to schools in a compensatory manner could best facilitate such accommodation. Schools were neither causing nor overcoming inequities students brought with them from home. By allocating effective school resources and varying the intensity of those resources proportionate to the varying outside/home resources, schools could create the most equitable educational opportunities for their students (Coleman, 1967).

While Sorensen and Hallinan (1986) found ability grouping had no impact on student’s individual reading achievement or growth, they also discovered ability-grouped classrooms were
the only place race played a significant role. In other words, when controlling for the impact of other variables, the effect of race on student achievement became a non-factor unless ability grouping had been employed. Specifically, Black students were both negatively affected by ability grouping and more frequently in schools that utilized the practice (Sorensen & Hallinan, 1986). Similarly, arguments have been made for detracking schools in the name of both equity and excellence (Oakes & Wells, 1998; Wheelock, 1992). Studies showed lower streams (tracks) led to lower academic demands and less engaging instructional techniques, culminating in students with negative views of their abilities. Furthermore, in streamed (tracked) schools, as in the United States, minority and low SES students were disproportionately represented in lower streams (tracks), reproducing existing social inequities (Smyth & McCoy, 2001).

Ansalone (2003) acknowledged tracking was used with the intent of increasing student achievement by creating classes of students more similar in ability, thereby creating more consistent targets for teachers. Nonetheless, he found tracking did not actually serve this purpose but instead took achievement gaps already present and widened them (Ansalone, 2003). Tracks, for a multitude of reasons, happened along socioeconomic and racial lines, and level of education implicated future profession. Together, a cycle perpetuated with a propensity to keep people in similar circumstances throughout generations (Ansalone, 2003).

Ability grouping began as a way to separate immigrants and Blacks from the apparently more educable Whites (Biafora & Ansalone, 2008). That ability grouping resurged around the same time of desegregation was not a coincidence but rather a systematic response to maintain the status quo (Chayt, 2010). Interestingly, in Ireland, tracking’s parallel practice of streaming had declined in the latter part of the twentieth century yet continued to be a common practice in schools serving largely disadvantaged populations (Smyth & McCoy, 2001). Recognizing the
negative effects of tracking on students in the lower tracks or streams, who were typically minority and disadvantaged students, an inequitable resource allocation practice was occurring. The very students experiencing the most detriment from tracking tended to be enrolled in schools that utilized the practice and, furthermore, were allocated to the lowest, most injurious tracks (Oakes & Wells, 1998; Wheelock, 1992).

Controversies Related to Grouping Practices in Schools

Grouping practices intertwined with segregation and its legal implications. Clearly, tracking is rooted in controversy that persists today, namely the divide between races and socioeconomic classes. *Plessy v Ferguson* (1896) could have marked the start of a sea change with respect to minority rights but instead reinforced Jim Crow laws and White-Black bifurcation. The ruling negatively affected education not only in New Orleans, where the case originated, but also across the country. In fact, it reinforced the denial of social and educational rights of Black people (Morris & Monroe, 2009). Over 50 years later, when *Brown v. Board of Education* (1954) entered the judicial system, the courts were ready to utilize the Fourteenth Amendment of the Constitution to extend equal protection of educational opportunities to all people, rendering “separate but equal” unconstitutional (U.S. Const, amend. XIV, § 1).

Not surprisingly, given the known segregative effect of tracking, the practice was also challenged in court, obtaining another ruling stating that “separate but equal” was unconstitutional. Chayt (2010) cited *Hobson v. Hansen* (1970) as a seminal tracking case in which the courts found the resulting segregation of Washington, D.C.’s tracking system to be discriminatory. While district officials argued the tracks were flexible, the court found the practice to be a proxy for segregation, which had only recently been addressed by the school district. Additionally, statistical analysis revealed almost no opportunity for students to change
tracks, with minority and low SES students disproportionately in the lower tracks (Chayt, 2010). Specifically, the court held “ability grouping <tracking> is by definition a classification intended to discriminate among students, the basis of that discrimination being a student’s capacity to learn” (*Hobson v. Hansen*, 1970). Interestingly, instead of initiating the end of tracking, similar cases brought to the courts after *Hobson* (1970) have not been found unconstitutional. In that *Hobson* (1970) had been such an egregious case, other cases failed to meet a similar standard of proof.

*Berkelman v. San Francisco Unified School District* (1974) challenged the constitutionality of an elite public school’s racial and gender admissions policies. The case raised questions of whether achievement scores were acceptable requirements for admission if the result was racial imbalance, whether having different male and female admissions requirements was permissible in order to maintain gender balance, and if having such an application-based public school was constitutional at all (Chayt, 2010). The court’s decision permitted the continuance of the race component in that it assisted a more diverse admission but disallowed the variable gender requirement, citing the Equal Protection Clause in the Fourteenth Amendment (U.S. Const, amend. XIV, § 1). The judicial system has long held education decisions were best left to local controls and educational experts. Part of this distance was reinforced by the *Plyler v. Doe* (1982) decision, which rendered education as close to a fundamental right as it likely will come, without actually declaring it as such.

Welner and Oakes (1996) described a growing awareness of the harmful effects of tracking. They indicated both researchers and reform advocates needed to maintain attention on the practice and consider legal bases for reform movements to ensue. While not ideal, court mandates to detrack, or some other kind of legislative action, could serve as effective policy
tools in particularly resistant areas of the country (Welner & Oakes, 1996). Chayt (2010) posited
the longstanding assumptions that benefits of ability grouping outweighed any negative impacts
contributed to an absence of litigation following Berkelman (1974). With the advent of NCLB
and disaggregated data reporting, perhaps additional court cases will ensue.

**Challenges with detracking.** As the desire to measure the effectiveness of educational
institutions and individual educators continues to increase, researchers, economists, and
statisticians continue to develop and refine methodological approaches and measurement
instruments. Should it be possible to have better, more reliable metrics and should those metrics
consistently point to the need to detrack public schools, it would be important to both know and
consider some of the challenges engaging in such a process would entail.

Following Slavin’s (1990) review of secondary tracking in which he found an average
achievement effect of zero, even the National Education Association (NEA) made formal
declarations against tracking as a practice, recognizing racial disparities across tracks (Argys et
al., 1996). However, after employing a standard education production function to control for
myriad variables while examining student achievement and its relationship to grouping practices,
Argys et al. (1996) indicated detracking may not be as much of a solution as was hoped. Others
offered that by focusing reform efforts on both the establishment of coherent standards with high
quality common assessments and the support of innovative, high leverage teaching, strong
negative reactions to perceived race-based change could be avoided or at least mitigated (Rubin

Relatively, Hochschild (1997) wondered if school desegregation was still a viable policy
option and argued that seeking to bring greater equity for all races by simply busing students did
not work well. Just like other reform efforts, busing alone would address a singular part of
education without addressing the larger school and organization. Rather, desegregation should be the result of, not the reason for, making other systematic changes within education (Hochschild, 1997). Such changes should be grounded within common goals as opposed to stemming from known politically or ideologically polarizing issues. Leveraging initiatives within the larger picture of reform offered a more palatable approach. Reaching out to the community for input, having clear goals, offering some choice within the efforts, and exercising morality-based leadership all could help lead the way to change (Hochschild, 1997). Ansalone (2003) agreed, cautioning if a school chose to undertake the task of detracking, there would be inevitable struggles, as deeply held beliefs regarding intelligence and ability would come to the forefront. Such a reaction could be tempered by carefully designed implementation and by taking time to address concerns of various stakeholders (Ansalone, 2003).

Similar to Hochschild’s (1997) recommendations, Rubin and Noguera (2004) suggested if a school decided to detrack, such a change would be best employed within a larger movement of school reform. Even if a primary goal of detracking were to lessen apparent segregation within a school, presenting the reform as a way to provide enhanced learning opportunities for all children would be a preferable message (Rubin & Noguera, 2004). In order to do so successfully from an instructional standpoint, substantial teacher training would also need to be part of the implementation plan. Otherwise, teachers may teach to the middle of the class, thereby leaving the highest level students and parents dissatisfied by the change in structure in addition to compromising some of the benefit to the middle and lower level students afforded by access and exposure to higher level material (Rubin & Noguera, 2004).
Student Achievement and Teacher and School Accountability

**Historical background of educational accountability.** To contextualize the current state of education reform, some historical background is needed. Specifically, the institution of education has been notoriously change resistant, yet recently significant changes around educator evaluation policy have been enacted. The motivations for these changes will be addressed. Notably, even though the responsibility for education was relegated to the states by way of the Tenth Amendment, over time the federal government has become increasingly involved in education (U.S. Const, amend. X). Indeed, though there have been calls for holding teachers accountable for their students’ performance since the mid-twentieth century, only within the last five years have many states begun to include student performance as part of their formal teacher and administrator evaluation systems.

The federal government has consistently placed a high value on education, with the courts stopping just shy of proclaiming education a fundamental right, citing the importance of an educated citizenry for the maintenance and forward progress of the nation. Historically, education has become a national interest largely in times of real or perceived crisis. When threats to America’s prowess arise, such as Russia’s 1957 launch of Sputnik or students’ consistently low rankings on international achievement tests, political leaders and citizens have repeatedly linked their fears to education (Darling-Hammond, 2010; Vinovskis, 1999). Following Sputnik, a sequence of events occurred to condition the present intertwining of education and the federal government. Under President Johnson’s Administration, the Elementary and Secondary Education Act of 1965 (ESEA) was signed into law, marking Congress’ first adoption of a general educational funding program. Such funding can be legitimized by way of the General Welfare Clause (U.S. Const, art. I, § 8).
While both this clause and the ESEA continue to serve as levers for federal influences on education, following Sputnik, the next major occurrence to bring education back into the national spotlight was the publication of *A Nation at Risk* (National Commission on Excellence in Education [NCEE], 1983). The report presented a dire picture of America’s public schools and America’s overall place in the world with respect to education. The tone of the report was alarmist, describing the educational system as “a rising tide of mediocrity that threatens our very future” (NCEE, 1983). In 1984, when President Reagan named four national education goals, it marked the first time a United States president declared goals related to education. Following suit, in 1985 the National Governors Association (NGA) offered to their respective states less regulation and control over schools in return for the willingness of school districts, leaders, and teachers to be held accountable for student progress toward well-articulated goals (Vinovskis, 1999). The bartering between accountability and state autonomy had begun.

Sitting presidents since Ronald Reagan have maintained education as part of their agenda. President Bush worked jointly with the nation’s governors to establish six national education goals at the Charlottesville Education Summit. Under the Clinton Administration, ESEA was re-authorized as the Improving America’s Schools Act of 1994. A notable change in the Act was the inclusion of a focus on student performance. Following the distribution of targeted funding, schools would be held accountable for having the achievement results of lower income students match that of other students. President George W. Bush continued the federal focus on education reform with the reauthorization of ESEA as the No Child Left Behind Act of 2001 (NCLB), formally signed into law in 2002 with high levels of bipartisan support. Notable changes in this Act were requirements, as opposed to suggestions, to establish academic standards and to institute testing for all students in order to benefit from supplemental federal
funding. *Sanctions* would also ensue for schools failing to make what was defined to be Adequate Yearly Progress (AYP), including a provision to “replace the school staff who are relevant to the failure to make AYP” (No Child Left Behind Act of 2001, 2002).

Furthermore, when teacher evaluation results have been studied, an overwhelming percent of teachers were given top or near-top ratings by their administrators (Whitehurst, et al., 2010; Donaldson, 2009). The preponderance of “excellent” teachers juxtaposed against America’s slipping ranking in international educational attainment coupled with persistent achievement gaps have added to the calls for more objective manners by which to assess teacher effectiveness, namely using student standardized test results as the metric. The stage was set for President Barack Obama’s Administration’s Blueprint for Reform, which included the Race to the Top Fund (RTTT). To receive funding, states had to hit specified policy targets, including the use of student testing data as a factor in the promotion and retention decisions of teachers (Teixeira de Sousa, 2010). While applying for the additional funding was optional, budgetary constraints led all but four states to apply for RTTT awards, which is almost singularly how the long called for policy changes of educator accountability have been swiftly enacted. Relevant to this study, in its RTTT application, North Carolina added an evaluation standard for both teachers and principals, which included for the first time consideration of student achievement as an indicator of educator effectiveness (Perdue, 2010). If student achievement results are related to classroom assignment practices, then attributing student achievement to teachers as part of their formal evaluation becomes problematic.

**Student Achievement Via Growth/Value-Added Metrics**

*Note from the researcher.* Initially, this study was going to utilize predicted and projected student performance values developed from EVAAS. In North Carolina, these values
are used in conjunction with students’ actual performance data to determine a portion of teachers’ and administrators’ official annual evaluations. While both the University’s Institutional Review Board and the dissertation committee had approved the study, obtaining individual student predictions and projections was not feasible due to the sensitive, personnel-based nature of EVAAS data. Therefore, the focus of the study no longer specifically involved Value-Added Models (VAMs) aside from the fact they are being used within accountability structures possibly affected by findings from the study. The statistical analysis employed in this study was not as sophisticated as a VAM but utilized similarly founded controls. The following abbreviated review of VAMs was included for reference and to further contextualize the issues around current accountability systems that are including student achievement metrics.

**Definition of growth/value-added.** In recent years there has been a shift from measuring student achievement via proficiency rates alone to also considering the rate at which students achieve growth, or one year’s worth of expected knowledge acquisition. Researchers suggested part of this shift resulted from a growing understanding that “static average student performance measures are poor indicators of school performance and tend to reflect input characteristics…” (Sloane, Oloff-Lewis, & Kim, 2013, p.39). An increasingly popular way to quantify such student growth is by utilizing value-added models (VAMs). VAMs use extensive historical data sets to predict or project a student’s performance on a state assessment. How well each student’s score aligns with the prediction or projection essentially determined whether a student hit his/her growth target. Confusingly, “growth” simply means a student learned what they would be expected to learn in one year’s time with an average instructor. In other words, students do not have to improve in performance from year to year in order to grow—they simply have to
maintain their learning position relative to others and at about the same level as they have performed previously.

In addition to considering growth/value-added metrics to assess student achievement, there has been a coinciding push to include student performance data in teachers’ evaluations. While teacher evaluation is not a new idea in education, actually utilizing student performance data for high stakes personnel decisions is relatively new and is being instituted at increasing rates across the country. The term value-added often refers to how much value a particular teacher adds to student learning. That is, do students overall tend to perform as well as, better than, or less than expected in relation to their predicted/projected growth targets while under the instruction of a particular teacher.

**Types of value-added models.** There are many different kinds of value-added models all purporting to accomplish the same thing: the gain score model, the covariate adjustment model, the Dallas value-added assessment system (DVAAS), the layered model including Educational Value Added Assessment Systems (EVAMS), the cross-classified model, the persistence model, and the Todd and Wolpin cumulative within-child mixed-effects model, along with others (Sloane et al., 2013). Notable differences include the inclusion or omission of student- and school-level context variables and whether or not the estimated teacher effects persist in students’ subsequent years’ achievement. Assorted combinations of these differences and how they are statistically utilized appear in the various models (Sloane et al., 2013). Which VAM model is applied can greatly alter the rating of a teacher’s effectiveness with the same group of students, bringing about possible rewards and sanctions in a seemingly arbitrary fashion. For instance, consider the variable of the persistence of a teacher’s effect on a student’s subsequent achievement. Some models include prior teachers’ effects at diminishing rates, while others
maintain each teacher’s full effect as part of each student’s future achievement. As such, a teacher’s current effect could be under- or over-estimated depending on both the manner in which prior teachers’ effects were handled and whether those prior effects were positive or negative (Sloane et al., 2013). While economists and statisticians continue to refine and develop new models, the rush to reform resulting from Race to the Top along with the marketing approach of for-profit SAS has led to the vast majority of the VAMs actually being used to evaluate teachers and administrators as EVAMS.

**Growth and VAMs in North Carolina.** Following the reauthorization of ESEA as the No Child Left Behind Act of 2001 (NCLB, 2002), a national movement toward state-level testing was under way. North Carolina had already instituted its own annual testing program and accountability efforts through the ABC’s of Public Education, first implemented in middle schools in 1996-97 and in high schools during the 1997-98 school year (North Carolina Department of Public Instruction, 1999). While North Carolina had been administering its own annual tests for particular areas since 1986, the ABC’s marked the first time sanctions and rewards would be tied to a school’s testing results. The ABC’s considered both achievement/proficiency and progress/growth rates utilizing NC’s own formula to develop growth targets. A primary benefit to examining student achievement data by way of growth attainment was the recognition that groups of students were inherently different. As such, cohorts of students were followed over time to detect growth, a more sensible approach than looking at how, for example, students performed in grade 8 one year compared to the previous year’s grade 8 students. Only certain subjects at each school level were tested, yet the overall results counted for the whole school’s staff. At qualifying schools, monetary rewards were distributed to every staff member, whether they taught a tested subject or not, at varying levels to incentivize higher
school performances in both proficiency and growth. Similarly, schools that did not perform well received “supportive” sanctions, which consisted of the deployment of assistance teams to those schools. While North Carolina employed their own growth formulas for fifteen years, they adopted the more sophisticated EVAAS system, a type of EVAMS run by SAS, as their growth model beginning with the 2012-13 school year.

**Concerns and implications of VAM use for accountability measures.** As including student achievement scores in personnel evaluation systems continues to make headway via Race to the Top grants, new concerns arise (Teixeira de Sousa, 2010). Similar to those cautioning against a focus on detracking in favor of comprehensive educational reform of which detracking is but one component (Ansalone, 2003; Hochschild, 1997; Rubin & Noguera, 2004; Smith & O’Day, 1991), Braun (2005) and others (see, for example, Corcoran, 2010; Donaldson, 2009; Goe, Bell, & Little, 2008) urged VAM metric use to be one of multiple ways to assess teacher effectiveness. He contended VAM metrics should not be used alone to make high stakes decisions, such as salary level or job status, as there were concerns about possible flaws in VAM application. He argued to determine a causal relationship, random assignment of subjects (students) to treatments (teachers) would be of fundamental importance. VAM use presumed such causality, yet students and teachers were not randomly assigned to one another, raising concern over the validity of VAM results (Braun, 2005; Rothstein, 2010; Sloane et al., 2013). While some VAMs sophisticatedly controlled for myriad variables, they could not substitute for randomization (Braun, 2005). Sloane et al. (2013) asserted “no set of adjustments can fully compensate for the lack of randomization” (p. 64). Of note for this study, the practice of tracking students into classes with similar ability students is decidedly not random.
While both reasonable and desirable to evaluate teachers according to their impact on student learning, the absence of causal assumptions and inherent bias on standardized tests (English & Steffy, 2011; Jenks, 1998) presents dilemmas for the use of VAMs as the only metric to assess a teacher’s effectiveness (Braun, 2005). The VAM used in North Carolina does not include any school or student contextual information in its modeling process. The model creators assumed such variables were captured in the student’s prior years of testing results (Ballou et al., 2004). Interestingly, EVAMS developers conceded the contextual capture assumption might hold less in places with greater stratification of teachers and students (Ballou et al., 2004). Indeed tracking exacerbated any existing strata, often not only in students but also in teachers, as their assignments to classes were also deliberate. Furthermore, a multitude of other variables have not been sufficiently studied in relation to their potential relationship with VAM metrics, including student-to-classroom assignment practices.

Of particular interest to this study, Koedel and Betts (2011) provided additional analysis regarding Rothstein’s (2010) critique of VAMs. Specifically, they investigated the extent to which VAMs were affected by unaccounted for variables, such as specific sorting bias. Such bias emanated from the manner in which students were assigned to teachers and classrooms for instruction (Koedel & Betts, 2011). Through their study, they found significant bias in teacher effect estimations that could be attributed to the sorting of students. However, they also showed that by utilizing complex VAMs, teacher effect estimates averaged over several years, the effect of the bias was essentially mitigated (Koedel & Betts, 2011). Even so, Koedel and Betts (2011) posited that in high stakes cases, it would be possible for deliberate student to teacher assignments to position teachers for differential results. They recommended safeguarding against such manipulation to protect both teachers and administrators. With respect to the
recommendation to only count a teacher’s VAM after several years’ worth of data could be collected and averaged, North Carolina’s recently implemented evaluation standard that is based on teacher VAM estimates becomes actionable only after a three year rolling average.

Policies often have unintended consequences. Such has been the case with the increase in school and teacher accountability, which has led to cheating on high stakes tests among other concerns (Amrein-Beardsley, Berliner, & Rideau, 2010). While VAM metrics measured different outcomes, the pressure on teachers and administrators nonetheless persists, especially in North Carolina and other states now including VAM results as part of personnel evaluations. In an ethnographic case study to examine possible NCLB impacts, Watanabe (2008) found substantive instructional differences in five areas, each favoring students in the “gifted” track. In agreement with other research, the class composition of the two gifted classes had significantly more White than Black students and the regular classes had significantly more Black than White students than reflected in the overall school composition (Watanabe, 2008). While intended to provide more equitable educational opportunities, movements such as No Child Left Behind have actually solidified differential instructional practices across tracks, with more explicit skill and test preparation instruction occurring in the lower tracks and more creative, engaging, and cognitively demanding instructional practices deployed in the upper tracks (Watanabe, 2008). Indeed, the residual effects of NCLB-spawned performance based accountability were differentially experienced across schools. Either under sanctions or threatened to be, lower performing schools, too often populated by minority and lower socioeconomic status students, focused more on test preparation, thereby narrowing the curriculum (Cohen-Vogel & Rutledge, 2009). NCLB has effectively led to an increase in tracking students by ability for instruction (Oakes, 2008).
Again, the more recent shift to a focus on growth and value added methods of assessing teacher effectiveness based on student achievement seemed promising. Yet their reliability has already been called into question. One example of this problematic dependability can be seen in teachers who had drastically different VAM results across classes in the same year and/or from one year to the next (Donaldson, 2009). Part of this differing in what seemingly should be similar results may stem from an assumption by value-added methods that tests were vertically scaled; that is, subsequent courses and tests in the same general subject would be more or less the same, only harder. This assumption was challenged by the reality of substantial content, not just difficulty level, and curricular changes year to year (Doran & Fleischman, 2005). Additionally, current value-added methods lacked future predictive capabilities. For instance, only a third of teachers rated in the top quartile of effectiveness were again in the top quartile the following year, while those rated in the lowest quartile had a ten percent chance of being in the upper quartile (Whitehurst, et al., 2010). Knowing this lack of stability, proponents suggested only considering VAM results for teachers after they have had three years of results (Sloane et al., 2013). It would seem to legitimately apply these methods for high stakes decision making, such as teacher or principal retention or promotion, the instrument would need to be proven more reliable, valid, and consistent. Yet, in order to receive federal monies, states, including North Carolina, have jumped ahead to change their evaluation polices.

While North Carolina had been reporting scores in the aggregate and using them to essentially rate, reward, and sanction schools for years, 2012 marked the first year in which every North Carolina teacher of a state-tested subject, along with school administrators, received a value-added model (VAM) rating as part of their formal evaluation. VAM ratings remain controversial, particularly in their application to individual classroom teachers. Rothstein (2010)
investigated the validity of the assumptions behind three VAMs, running various falsification tests against them. He found a teacher’s current year impact on student achievement correlated only between 0.3 and 0.5 with their two-year impact (Rothstein, 2010). Rothstein (2010) focused on within school differences due to the fact students were not randomly assigned to schools. He remained concerned variables unable to be captured and controlled in VAMs would lead to inappropriate rewards and sanctions delivered to individual teachers.

Dismantling some of the VAM credibility, he found significant relationships and predictability between fourth grade teachers and their students’ fifth grade gains. Even more quizzically, a student’s current year teacher was found to have a significant impact on their prior year’s gains (Rothstein, 2010). Ultimately, Rothstein (2010) established the assumptions behind all three models were incorrect, raising concerns about their validity, at least in North Carolina, where he completed this study. Rothstein (2010) suggested several ways to make the VAM more reliable, including the addition of some school level class assignment process information, indicating how students were assigned to classes may be an explanatory variable. Rothstein (2010) asserted without “high-quality principals who have enough time to observe teachers’ classrooms and enough training to distinguish good from bad teachers…neither subjective nor VAM-based estimates that depend importantly on classroom assignments are likely to provide much useful information” (p 212).

On the other hand, Koedel and Betts (2011) demonstrated by utilizing particular types of VAMs, those with a complexity level commensurate with the teaching–learning dynamic, the sorting bias found by Rothstein (2010) diminished. Indeed, even with less complicated VAMs, should a multi-year average VAM effect be used to rate teachers, previously seen student sorting bias disappeared to statistically insignificant levels (Koedel & Betts, 2011). Nonetheless, Koedel
and Betts (2011) cautioned administrators could affect teacher VAMs, should they wish, which necessitated sufficient safeguards against such practices. Furthermore, their study was based on low stakes assessments. High stakes assessments substantially altered the climate for VAMs, again pointing to the need for a system of checks and balances regarding how student class assignment decisions were made. A primary implication of their findings was the need to utilize multiple years of data to sufficiently account for possible sorting bias and result in a fair estimate of a teacher’s effectiveness (Koedel & Betts, 2011).

**Gaps in the Literature**

How is this discussion of accountability and VAMs relevant to student-to-classroom assignment practices? Rothstein (2010) pointed out the use of VAM metrics to essentially rate a teacher’s effectiveness assumed the model supported causality. However, at a minimum, a study must include random assignment of subjects to treatments for causality to be determined. In this study, students would need to be randomly assigned to teachers (Braun, 2005; Rothstein, 2010; Sloane et al., 2013). Unmistakably, if students were being assigned to classes based on ability, randomness was not possible. Even if it were possible to randomly track students by ability, there would remain issues of favoritism, separation of specific students, along with different levels of classes. Proponents of VAMs believe the models account for any and all variables, thereby negating unfair advantages/disadvantages a particular teacher may have with any group of students—this is largely the rationale behind utilizing a growth-based teacher effect model. However, if the kinds and levels of students a teacher is assigned to teach have an impact on the teacher’s propensity to obtain a high or low VAM rating, then using said ratings for teacher evaluation would border on being arbitrary and capricious, setting up inevitable litigation. Furthermore, principals may intentionally position teachers for differentiated outcomes
depending on factors not always aligned with the primary work of teachers. How student-to-
classroom assignment practices relate to relative student achievement and VAM metrics must be
investigated to both prevent such power abuses and to re-examine student learning across
different organizational practices through a growth and/or VAM lens.

**Literature Themes and Critique**

Acknowledging the role context plays within the human organizations of education, it is
no surprise there is not a one size fits all answer to the call for best practice related to grouping
students for instruction. Equally expected, there are conflicting study results and research
findings. Nonetheless, some common themes emerged among the literature. Multiple studies
concluded high-level students benefited from either an academic or a social and emotional
standpoint when grouped with similar ability students (Adams-Byers, Whitsell, & Moon, 2004;
Argys et al., 1996; Conger, 2005; Hallinan, 1994; Lleras & Rangel, 2009; Neihart, 2007;
William & Barholomew, 2004). At the opposite end of the spectrum, research pointed to lower
level students tracked together for instruction experiencing a detrimental effect on their gains and
achievement (Ansalone, 2003; Lleras & Rangel, 2009; Oakes & Wells, 1998; Smyth & McCoy,
2001; Sorensen & Hallinan, 1986; William & Bartholomew, 2004). The only study reporting
benefits of tracking for students at all levels came out of Kenya (Duflo et al., 2009). While the
Kenyan study found tracking served all students better than non-tracking, the learning within the
tracks was commensurate with the levels. In other words, students were not learning the same
material at the same depth. These contrasts, if overall true for most students, raised significant
ethical issues related to educational practice: do we track students by ability, knowing the impact
of doing so will further stratify the population into academic haves and have-nots? On the other
hand, do we knowingly take away the benefit of ability grouping from the high level students in
order to provide greater access and opportunity for the lower level students? Clearly these are questions without easy answers.

A commonly cited reason for concern, at a minimum, or a call for moral action, at a maximum, related to tracking was the appearance of segregation between upper and lower tracks (Ansalone, 2003; Ansalone & Ming, 2006; Oakes & Guiton, 1995; Rubin & Noguera, 2004). The numbers simply did not lie: there were a disproportionate number of minorities and lower socioeconomic students in the lower tracks throughout K-12 education and the discrepancy worsened in upper grades. The pernicious remnants of the ideologies from the 1920’s, as illustrated by the following quote from Lewis Terman (1923), resonate too closely with current realities: “Their lack of intelligence is racial, but while they cannot understand abstract concepts, let us make them efficient laborers” (p. 28). As Oakes and Guiton (1995) pointed out after drilling-down into the demographic disparity between tracks, the reasons for such misrepresentation were complex and deeply rooted within perceptions, curriculum, and structure. Families and students advantaged by innate ability, socioeconomic status, or support structure, maintained their advantage in multiple and structured ways while the opposite reality ensued for the disadvantaged. Remarkably, the same issues that drove tracking decisions long ago play into our current mindset with regard to the contended purpose and role tracking plays in our schools today (Biafora & Ansalone, 2008).

It would seem if public schools in America truly strived to be the equalizers of opportunities for all, then heterogeneously grouped classes would more likely serve that purpose. If schools were to choose this route, there is wide agreement significant professional development would be needed. With training, teachers could better match students’ learning styles and differentiate instruction within mixed level classes, thereby maintaining a similar level
of challenge for the upper level students (Ansalone & Ming, 2006; William & Bartholomew, 2004; Neihart, 2007; Rubin & Noguera, 2004). Indeed, researcher such as Ben-Ari and Kedem-Friedrich (1999) determined that, provided the teacher was skilled in facilitating student learning, mixed level classes would be best for all students’ cognitive development.

A final emerging theme observed in the tracking literature dealt with the great care and thoughtful intention needed to successfully move from a tracking-based school to a more equal-access based school. There were no easy answers to solve a given school’s problems (Archbald & Keleher, 2008; Hochschild, 1997; Rubin & Noguera, 2004). The context within which one attempted to develop such change played a critical role in its reception by both the community and staff. One must know intimately the details of the school, along with its culture and climate, and then be grounded firmly within a commonly agreed upon reason to act.

For myriad reasons, the composition of schools remained an important factor for parents, often impacting where families decided to live within a county or a state (Hanushek et al., 2003). Neighborhoods and their receiving schools were affected by poverty and wealth, factors outside, yet inextricably linked, to the system (Coleman et al., 1966). Some believed poverty was used as too much of a scapegoat for educational woes (Ansalone & Ming, 2006), while others contended without addressing poverty, true change in educational outcomes was unlikely (Berliner, 2006). Ansalone and Ming (2006) suggested there was a tendency to blame students’ low levels of achievement on their circumstances, leaving schools too much out of the picture. They asserted too many of the lower achieving, and therefore deemed lower ability, students were from poor families. Schools needed to increase efforts to address student-to-student differentials observed upon school entry (Ansalone & Ming, 2006; Coleman et al., 1966). Berliner (2006), on the other hand, pointed out students spent five times the number of hours outside of school as within,
suggesting the need to address systematic inequities outside the classroom. Doing so could impact social and cultural capital acquisition, invariably affecting student achievement. Berliner (2006) argued to truly shift a community, each component of within-school reform movements—teacher qualifications and expectations, for example—needed to coincide with an equally impactful outside of school reform movement, such as livable wages and affordable healthcare.

Ever since the issuing of *A Nation at Risk* (NCEE, 1983), there has been a ratcheting up of desire to hold educators accountable for the outcomes of their work, namely the learning of their students (Veir & Dagley, 2002). As the demands placed upon education from the federal level have increased, the discussion and rhetoric around the idea of heightened accountability within education has shifted as well. There has been a progression from a focus on inputs into the system, i.e., the credentials and preparation of teachers and the structural requirements for high school graduation, to a focus on outputs, namely student achievement as measured by graduation rates and performance on standardized tests. The latest accountability shift has been twofold: include student achievement results in teacher and administrator evaluations and utilize VAMs to do so. Derived from the RttTF, the unprecedented budget to reward states taking swift action to institute these accountability shifts and other reform efforts has led to hasty implementation, seemingly without much consideration of consequences (Teixeira de Sousa, 2010).

If advanced statistical analyses, including VAM metrics, were reliable and valid and truly able to determine how much “value” a teacher added to their students’ learning trajectories, they would hold promise for use as one of multiple measures of teacher effectiveness. Yet there are concerns regarding their violation of randomization assumptions, questions about their ability to account for context, and myriad models from which to choose (Braun, 2005; Donaldson, 2009;
Paufler & Amrein-Beardsley, 2014; Sloane et al., 2013). The lack of stability in teacher VAM results within and across years along with studies indicating a mere thirteen percent of the variance in student achievement was attributable to their teacher adds to the concern over using statistical modeling related to student achievement metrics for personnel evaluation purposes (Cissell, 2010; Donaldson, 2009; Whitehurst et al., 2010).

Even critics of NCLB, such as Consiglio (2009), conceded the reason the federal government has been stepping into the state matter of education stemmed from the reality that prior recommendations had not altered outcomes or changed practice. Schools were notorious for being change resistant (Lortie, 1975; Ogawa, 2009; Wagner, 2008). There continued to be a need for educational reform, though some asserted the focus on testing and accountability has rung hollow with educators and would not stand alone as effective reform. Rather, tandem attention should be given to instructional and curricular reform—instilling one without the other would result in limited success or change (Consiglio, 2009; Ravitch, 2010; Smith & O’Day, 1991). Tracking, as it has become a relatively unquestioned practice, must again be scrutinized. Hegemony develops when the consistent power and influence of the dominant group becomes the norm, accepted as “the way things are”—tracking has become an accepted part of the public school fabric, too often escaping interrogation. Critical theory troubles such mindless or intentional maintenance of the status quo and dominance. In this way, critical theory can offer needed perspective with regard to framing complex educational issues by utilizing intentional oppositional thinking to present issues and ideas plaguing society and school systems (Barbour, 2011). Instead of avoiding conflict in favor of a uni-dimensional and hegemonic approach, critical theorists uncover conflict as a way to bring about change and action, giving voice to traditionally marginalized groups. Multi-dimensional approaches to solving the layered problems
are needed—solutions likely must involve imagination and creativity to break free from patterns of the past. Certainly tracking is an issue with many layers, both historically and culturally, and its relationship with student achievement as rendered by newly developed analytics metrics is underexplored.

**Chapter Summary**

Since its inception, the arguments for and against tracking students into classrooms by ability have largely remained the same (Betts & Shkolnik, 2000; Kulik & Kulik, 1982; Slavin, 1990). Proponents based their arguments on the belief ability grouping was the most effective and efficient way to teach students, whereas opponents pointed to inequity as the primary reason to not institute ability grouping—such inequity arose from the varied experiences afforded students in different tracks and the disproportionate representation of minority students in lower tracks. Furthermore, tracking proponents tended to be more outcomes-oriented while opponents included a focus on values, such as social justice and racial equity. As such, the “burden of proof,” so to speak, was on the proponents, as their primary argument was based upon ability grouping’s effectiveness and outcomes. Opponents, on the other hand, presented compelling arguments to dismantle the practice of tracking students by ability even if it were effective with respect to outcomes. They have argued ability grouping perpetuated the status quo, which favored hegemony and made escape from lower socioeconomic strata extremely challenging. Tracking, therefore, was antithetical to democratic ideals. With the advent of performance-based accountability for both teachers and schools, a new element must be considered: if VAM metrics and other measures of relative student achievement are a better way to assess student growth and teacher effectiveness, what is the relationship between tracking and relative student
achievement? This examination is urgently needed as teachers and administrators are already receiving VAM ratings in their official evaluations.

If student-to-classroom assignment practices interact with relative student achievement or VAM metrics, then teacher effectiveness ratings based on VAMs either need to be discarded or statistically restructured. It is possible the rush to federal monies will result in an unwarranted rush to judgment that impacts educator livelihoods. A somewhat more familiar aspect to the study is the examination into how assignment practices affect student learning. Of primary concern is whether or not differential relationships with student learning by level and by classroom assignment practice exist. If they do, important equity issues will surface.

This study seeks to address a gap in the literature to connect student-to-classroom assignment practices with relative student achievement metrics. The following chapter will present the manner in which these questions will be addressed.
CHAPTER THREE: METHODOLOGY

Introduction

This chapter discusses the rationale behind the research design employed to study the relationship between student-to-classroom assignment and relative student achievement metrics from standardized state tests. The methodology and procedures for access, data collection, data organization, and data analysis are presented.

Purpose, Research Hypotheses and Rationale for Study

The extant literature about how tracking may or may not be related to student achievement is substantial (see, for example, Ansalone, 2003; Betts & Shkolnik, 2000; Oakes, 2005; Slavin, 1990). Continuing with the premise that a relationship may exist, classroom assignment presented the first line of demarcation for the study. Likewise, the culmination of the study remained performance-based, albeit a more relativist performance metric, taking into account a student’s prior performance along with other variables. The decision to disaggregate student relative achievement was based both in the literature and in the No Child Left Behind Act of 2001. The literature indicated discrepancies in student achievement between levels of students in addition to racial disparity both in proportionality within levels and in performance (see, for example, Ansalone, 2003; Lleras & Rangel, 2009; Oakes, 2005). Disaggregating data has become more commonplace since the issuing of NCLB, which required and inspected the reporting of performance data by various subgroups.
Tracking has a controversial history, falling in and out of favor in cyclical ways often corresponding to the political climate. Prior to the onslaught of standardized tests resulting from NCLB, tracking had become less prevalent. School leaders and politicians recognized a need to have a more egalitarian approach to education. This change precipitated from advancements in civil rights as well as a shift in the preparation needed to either join the workforce or attend college. However, as testing results began to be reported, particularly in the NCLB disaggregated manner, leaders retreated to prior tracking methods, emboldened by the indicated disparate outcomes. It seemed as if administrators saw achievement gaps as proof positive that what they were currently practicing, a more mixed level approach to classroom assignment, was ineffective. They reasoned a return to sorting students by ability for instruction was prudent, and the resurgence of tracking had begun (Loveless, 2013). An underlying assumption, then, was that how students were assigned to classes had a relationship with their achievement. Was that the case? If so, what was the relationship?

The purpose of this study was to investigate the relationship between student-to-classroom assignment practices in the secondary school setting and relative student achievement on North Carolina English/Language Arts standardized state tests. The study’s major research hypothesis was that a relationship does in fact exist in ways that support the following sub-hypotheses:

**Sub-Hypothesis 1:** Non-advanced students have better relative achievement when they are assigned to mixed level classes than when they are in tracked classes.

**Sub-Hypothesis 2:** Advanced students’ relative achievement is independent of the type of class to which they are assigned.
Sub-Hypothesis 3: Mixed level classrooms better reflect the school’s racial distribution than tracked classrooms.

Conceptual Framework

Social Learning Theory

Bandura’s (1977) social learning theory asserted people learned primarily through the observation of others. In other words, people tended to become more similar to each other through incorporating behaviors modeled by others in both personal relationships and group settings. He stated, “…most human behavior is learned observationally through modeling; from observing others one forms an idea of how new behaviors are performed…” (p.22). Similarly, Vygotsky (1978) theorized the development of a child occurred socially, that is, between people. He posited all learning and development occurred in this dynamic fashion prior to being internalized by the child. Indeed, both theories supported the idea that it mattered beside whom a child learned. Presumably, in heterogeneously grouped classes, higher achieving students would positively affect lower achieving students. At the same time, the lower achieving students may adversely affect higher achieving students. Continuing this line of argument, tracking would be beneficial for higher achieving students and inconsequential or unfavorable for lower achieving students. As such, the study design essentially tested these theories of learning and development.

Critical Theory

Critical theory provided the framework to interrogate classroom assignment practices. As indicated in the literature, when students are assigned to classes by ability, African American and Latino students are disproportionately assigned to low tracks, while White students are disproportionately assigned to high tracks (Ansalone, 2003; Oakes & Guiton, 1995). As such, in that research has demonstrated both the quality of instruction and level of expectations are higher
in advanced track classes, White students tend to be exposed to a higher quality education. Gillborn (2008) asserted such structural differential access to quality education, even within the same school, was a manifestation of institutionalized racism. Reinforcing how ubiquitous grouping students by ability had become, Oakes (2005) referred to tracking as “one of those relatively unquestioned practices that belongs to the ‘natural’ order of schools” (p. 191-192). Critical theory questions such maintenance of the status quo, regardless of intention, because it contributes to the cultural capital of the elite class, perpetuating class division and providing a power base from which to operate.

In order to (re)develop more equitable education policies and systems, the effects of current and past policies on marginalized populations and their possible originations must be known. In part, this study sought to exploit the availability of achievement data from the same students under different grouping schemes. The intention was to examine whether lower performing students could do better given the assumed benefits of a mixed ability environment. Such a finding would discredit the response that students were grouped due to innate and prohibitive abilities. This study also sought to address the other common argument that advanced students’ progress would be hindered if they were taught in the same classroom with non-advanced students. Critical theory provided a framework for analysis with additional attention to race and can serve as an invaluable tool in the process of dismantling an entrenched practice such as tracking (Capper, 1993; Crenshaw, 1995; Dixson & Rousseau, 2005).

Application of the Framework

Essentially, then, the study was conceptualized as the actualization of both social learning theory and critical theory and their implications when students are assigned to classrooms. If, as social learning theory asserts, children learn as much from their surrounding peers as any other
factor, tracking would be beneficial to students in higher tracks and detrimental to lower track students. Such differentiated effects would replicate societal strata in ways critical theorists reject.

Based on this study’s purpose, hypotheses, site, and conceptualization, the study frame mapping with embedded conceptual framework components on the following page provided guidance for data collection and analysis. Available categorical data were classroom assignment practice, teacher class period, honors/advanced versus standard/non-advanced level enrollment, teacher combination, and other student demographic information. Quantitative data in the form of historical testing results for two or three consecutive English Language Arts (ELA) standardized tests comprised the primary data of interest in the study. Relative student achievement was determined for each subgroup by the differences between students’ average performance in sixth and seventh grades when they were tracked, and eighth grade when most students were in mixed ability classrooms. The ultimate question was whether or not there were differences in the three student groups’ relative achievement, thereby determining the equity or lack thereof resulting from classroom assignment practices. Tables and spreadsheets with the aforementioned categories and codes helped to organize the data. The framework also illuminated the kind of analyses required. Tests for differences in relative student achievement among the student-to-classroom assignment trajectories were needed. Testing for differences across levels but within the tracked-to-mixed trajectory provided information with regard to possible differential relationships, as did tests for differences across grouping trajectories but within the same level of student, namely the non-advanced students. Descriptive statistics, matched pairs tests, ANOVA, ANCOVA, Chi-square tests, and multilevel modeling all contributed to the final analysis.
Figure 1. The interaction of the guiding conceptual framework with the study design.
Rationale for a Quantitative Study

Based on the study’s major research hypothesis, the outcome of interest was relative student achievement on state standardized tests. Having a numerically measured continuous outcome variable necessitated a quantitative study. Furthermore, while myriad other variables warranting a qualitative study and analysis may be of interest, such as teacher or student perceptions of grouping practices, the focus of this study was only on classroom assignment practices and their possible relationship with relative student achievement on state standardized tests. Limiting the variable of interest allowed for more robust data collection and analysis. Relative student achievement, as determined via advanced statistical modeling, has been shown to be a more authentic measure of student learning or growth (Ballou et al., 2004). Though the legitimacy and appropriateness of such measures are debated, some states are already using such metrics to ascertain teacher effectiveness, at times used as part of evaluation and/or to determine compensation awards (Springer et al., 2010; Teixeira de Sousa, 2010). If classroom assignment practices have a relationship with either or both relative student achievement and/or aggregate student achievement linked to teachers, the research indicating this impact must be applicable to broad settings. Applying research to scale and obtaining results from which inferences can be made about general populations are benefits of quantitative studies (Creswell, 2012).

Site Selection and Participants

Access

To complete this study, the researcher needed access to existing data sets. Student performance data has become part of teacher evaluation in North Carolina, resulting in additional protections and access restrictions. Requesting data access from within the researcher’s current district of employment enhanced access likelihood. Additionally, many gate-keeping
administrators lack both familiarity and confidence with data comprehension and analysis. Bolstered by a background in teaching statistics, the researcher explained the research goals and offered the administration one-on-one sessions to review any data or reports. In an offer of reciprocity, if the administrators desired, the researcher proposed consolidated research summaries to accompany the results of the study (Marshall & Rossman, 2011). For the purposes of this study, the researcher secured access to and approval of the use of data from an ideal site, a school employing both kinds of student-to-classroom assignment techniques. Depending on the grade level and situation, the school sometimes tracked and other times heterogeneously grouped students for instruction. The purposeful selection of this school was a type of convenience sample, ultimately benefiting the researcher with access and the study with an ideal situation in which the same students experienced tracked classes one year and mixed classes the next year (Gall, Borg, & Gall, 1996).

Located within an urban school district, this magnet school offered a sample more racially and socioeconomically representative of the district than would a neighborhood-based school, which largely followed catchment lines reflecting socioeconomic and racial stratification. The school’s magnet program was whole school, and being selected via a blind lottery was the only way a student could enter the school. At the same time, sampling from a magnet school introduced a different kind of selection bias in that all students entering the blind lottery to join the school had parents or guardians interested and aware enough to have known about the school and to have filled out an application. While other researchers, such as Ballou, Goldring, and Liu (2006), have been able to add a control element for this unique form of bias by also following students who applied to but did not win the lottery, this researcher could not do so. The benefits of having both “treatments” applied to most students in the same school outweighed including
other schools’ students. Furthermore, selection bias was mitigated by the study design as students were compared with themselves and to students within the same school setting. The school served approximately 1600 students in grades six through twelve, with approximately 200 students in each of the middle grades. The school composition with respect to race was as follows: 38% African American, 34% White, 19% Hispanic, and 9% Other. Unlike other schools in the district, this school’s population fairly well mirrored the racial composition of the city it served: 38.5% African American, 42.5% White, 13.5% Hispanic, and 9.9% Other (United States Census [USC], 2013).

Data Collection

The Institutional Review Board at the University of North Carolina at Chapel Hill, the school district’s Research Review Committee, and the school principal all approved the study and the access required to complete it. In addition to access to the school’s student performance data, copies of prior master teaching schedules and class rosters were necessary. The researcher also extracted student demographic information from available databases.

Population and Sample Size

The population of interest in this study was North Carolina public school secondary English/Language Arts (ELA) students. Two performance levels of students, advanced and non-advanced, defined subpopulations of interest. For the purposes of this study, advanced students were those enrolled in the advanced level class sections and all others comprised the non-advanced students. The vast majority of advanced ELA students were in the advanced section as a result of being identified as Academically and Intellectually Gifted (AIG). Teachers recommended a few other students for the advanced level who had shown significant promise the prior year. Of note, the AIG label is derived differently across districts, and this district liberally
applied the label to essentially the top 20% of the district’s student population. At this school, approximately 37% of the students were identified as AIG. The entire cohort of roughly 200 students enrolled in 7th grade ELA (ELA7) during the 2011-2012 school year and in 8th grade ELA (ELA8) at the same school during the 2012-2013 school year comprised the sample for the study. This sample was considered representative of prior and forthcoming cohorts of students at the same school due to the demographic and achievement stability of the previous five student cohorts.

During the 2011-2012 school year, two different 7th grade four-teacher teams taught the students in this study, with approximately 100 students populating each team. On each team, each of the four teachers taught one subject: ELA, Math, Social Studies, or Science. On one team, ELA7 Teacher A taught one advanced, one non-advanced, and two non-advanced ELA sections that included Exceptional Children (EC) and inclusion services. On the other team, ELA7 Teacher B taught two advanced and two non-advanced sections of ELA. This same teaming structure had occurred with a different set of two four-teacher teams in 6th grade the year prior. However, after students experienced two tracked ELA years, the following year, 2012-2013, the same students were rearranged into six sections of heterogeneously grouped (mixed) ELA8 sections and, again, into two non-advanced sections with Exceptional Children and inclusion services. Students were assigned to either ELA8 Teacher C or ELA8 Teacher D, both of whom taught three mixed sections and one EC non-advanced section. In 8th grade, students attended core classes without the constraints of the traditional middle grades teaming structure. Figure 2 presents the change in grouping practices between 7th and 8th grades.
The number of students attending the school consecutively for grades seven and eight determined the size of the operational sample. Students in either grade who had not attended the school during the other grade were eliminated from the analysis. There were four subpopulations of interest to the study and applicable to other settings: advanced students who moved from tracked to mixed level classrooms, non-advanced students who moved from tracked to mixed level classrooms, advanced students who remained in tracked classrooms for consecutive years, and non-advanced students who remained in tracked classrooms for consecutive years. However, this study could only examine three of these four subpopulations. Due to the selected school’s student-to-classroom assignment practices, there were no advanced students who stayed tracked in both 7th and 8th grades.

**Rationale for Choice of Participants and Sample Size**

Beyond selecting an urban school district, accessibility played a role in both district and school choice. The researcher leveraged relationships and employment status to gain data access.
Urban districts were of interest due to myriad concerns around the practice of tracking, many of which were equity-based and more prevalent in urban settings (see, for example, Ansalone, 2003; Oakes, 2005). Site selection was made based upon the type of student-to-classroom assignment techniques utilized by each school. The researcher used knowledge of a school employing both tracked and mixed level classroom assignments, and so chose to first pursue access to that school’s data. As a result, the researcher inherently controlled for multiple variables, such as the culture of the staff and school, location, schedule, etc. For example, in that all individuals attended the same school at the same time and served as their own control, there was no need to control for a school effect.

Additionally, matching students to similar students was unnecessary, because students served as their own control, or matched pair, over time. Capitalizing on individual matched pairs offered additional control for possible confounding variables (Yates, Moore, & Starnes, 2002). While research has shown some principals use student characteristics, such as race, ability, and behavior as factors in assigning students to classes and others have reported senior or favored teachers receive preferential class assignments, such was not the case at this school (see, for example, Burns & Mason, 2002, and Kalogrides, Loeb, & Béteille, 2009). Students were assigned to classes via computer generation. Specifically, each 8th grade mixed class pulled in both levels of students, while the two tracked classes only pulled in non-advanced students and those who required inclusion services. Following the initial computer classroom assignments, the 8th grade teachers followed up to ensure as much balancing with respect to race and advanced/non-advanced students in the mixed classes. Lastly, both the teams of teachers and ELA teachers were stable across the years in question, and, notably, both pairs of ELA teachers planned instruction together and utilized the same resources and activities in their classrooms.
Given the study’s aim, the selection of ELA data was intentional for several reasons. First, ELA offered the distinct opportunity wherein the majority of students experienced a change in grouping practice from one year to the next, namely moving from a tracked class in 7th grade to a mixed class in 8th grade. Second, reading comprehension, which is the primary metric on the state ELA assessment, has been referenced as a construct of both inside and outside of school influences, while mathematics achievement has been shown to be more affected by direct schooling experiences. Therefore, should substantial performance changes from 7th to 8th grade ELA be discovered, one could argue the effect of grouping structure was even more significant than would be a parallel change in mathematics. Additionally, due to the advancement of many 8th grade students into high school level mathematics courses, an analogous change in grouping for nearly all students at the same time was not possible. Lastly, literacy skills, including reading, are required for success in many subjects, not just ELA. Therefore, improvement in ELA test performance could serve as a proxy to indicate similar gains in other reading comprehension dependent courses in the middle grades, namely science and social studies.

Research Design

Overall Design

This study was set up as an ex post facto correlational design. The researcher did not actively manipulate a treatment, thus making it an ex post facto study. Instead, the researcher observed how the treatment, student-to-classroom assignment practice, may be related to secondary students’ relative achievement within and across advanced and non-advanced levels of student. The study was correlational in that collected data were used with the primary goal of ascertaining “…whether, and to what degree, a relationship exists between two or more quantifiable variables” (Gay, Mills, & Airasian, 2009, p.196). In this study, student-to-
classroom-assignment trajectories being different within the same cohort of students within the same school offered a unique opportunity to gather evidence of possible relationships between how students were assigned to classes and their relative student achievement. In this case, examining longitudinal data for changes over time corresponded with a change in or maintenance of a student-to-classroom assignment practice. The design was a factorial design, with two types of classroom assignment techniques comprising three trajectories across 7th and 8th grades for advanced and non-advanced students (Creswell, 2012).

The study framework illustrated elements of a block design whereby in 7th grade, when all students were tracked, they were blocked by level for analysis. Students within 8th grade mixed level classrooms were also separated by level. Blocking controlled for the variable of overall level of student, which may be related to relative achievement. Creating “treatment” groups as homogenous as possible allowed better assessment of the possible relationship between the treatment and the outcome (Yates et al., 2002).

Variables of Interest

For implications of both equity matters and teacher evaluation results, the dependent variable analyzed was the relative student achievement from year-end North Carolina English/Language Arts standardized state tests. The North Carolina Department of Public Instruction (NCDPI) created, maintained, and delivered annual reports of student performance to North Carolina school districts. These reports were available within schools within districts, thereby providing the needed data for the principal dependent variable of interest, student relative achievement. The primary independent variable of interest was the type of student-to-classroom assignment practice used, namely tracked or mixed level assignment. A secondary independent variable of interest was the level of the student. Level of student, advanced or non-
advanced, related to prior student performance and was determined by level of ELA in which a student was enrolled, honors/advanced or standard/non-advanced. Theoretically, due to the matched pair nature of the student outcomes, time-invariant predictors could be removed from the model, at least from the individual student. The researcher examined independent variable effects both with and without such covariates.

Other Variables

While not the primary variables of interest, in settings such as schools, one must consider other variables that may contribute to achievement variation. This study included information to examine and account for possible peer effects and teacher effects (Hanushek et al., 2003; Horoi & Ost, 2014; McCaffrey, Sass, Lockwood, & Mihaly, 2009). Both cases attempted to address possible homophily; that is, the tendency for things, in this case students, to be more similar to each other when they were in the same classroom, peer effects, or taught by the same teacher, teacher effects (Grunspan, Wiggins, & Goodreau, 2014). In this study, 7th grade teacher effects could be proxies for overall team effects as students of a particular ELA teacher were also taught by a linked team of three other core area teachers. Such effects could be positive, negative, or negligible. Given the dynamic nature of the teaching-learning process, it was prudent to attempt to account for as much achievement variation as possible. In order to do so, the researcher needed information for each student, including to which teacher and class period they were assigned. This information was obtained from the school’s database, including class rosters and school master schedules.

Procedures

Following approval from the Institutional Review Board of the University of North Carolina and obtaining access to student achievement and demographic data from the school and
school district, the researcher acquired needed information from all available data sources. Prior to analysis, data were organized, coded, cleaned, and de-identified. More detailed procedures are discussed later in this section.

**Existing Data Sets**

Existing databases owned by the state and held within schools and school districts provided the information needed to complete this study. Due to the recent inclusion of student performance data in teacher evaluations, access to data sets has become more restricted. As such, the researcher needed to leverage personal connections to gain school-specific level access to the data of interest. To further facilitate access approval, the researcher presented as trustworthy, knowledgeable, and forthcoming regarding goals of the study (Marshall & Rossman, 2011). The fact that participation in the study involved no known risk also positively impacted data permissions. Additionally, a well thought out plan to maintain the anonymity and protection of student, teacher, school, and district information allowed district and school officials to feel more comfortable in granting access to the data (Creswell, 2012). The school district assigned a school level research site administrator who ensured appropriate data safeguards were followed. The researcher used pseudonyms at all times to protect the identities of any subjects, schools, or districts included in the study. To further protect student level information, the researcher assigned a random numerical code to replace each individual’s name or other identifying information. The research site administrator witnessed the de-identification of the data to be used in the study. The linking file was maintained until the final elements of the study were complete. The researcher worked with coded information from the point of acquisition forward. Furthermore, working data files were maintained on a secure, password-protected drive to which only the researcher had access.
Data Compatibility

The NCDPI reported standardized test results for each student using both a scale score and a percentile value. In order to use cross-year longitudinal data by student on North Carolina state English/Language Arts End of Grade (EOG) standardized tests, results were converted to Normal Curve Equivalents (NCEs). Normal Curve Equivalents offer a way of standardizing test scores to an equally scaled value between one and 99. NCEs have a mean of 50 and a standard deviation of 21.06 (“Stata FAQ”, n.d.). Percentiles indicate relative frequency distribution information and so do not follow an even distribution, whereas NCEs reference position on an equal-interval scale. Specifically, a difference or change of \( n \) points in NCEs communicates the same student performance information regardless of positioning within the distribution of outcomes. For example, a difference of five NCEs translates into the same amount of movement on the x-axis no matter whether the change occurs at the bottom, middle, or top of the distribution. A change of five percentile values, on the other hand, indicates substantial movement at the top and bottom of the performance spectrum but negligible movement if originating in the middle of the distribution. Therefore, by using NCEs, measurement bias more prevalent in the low and high ends of the achievement spectrum was minimized (Lockwood & McCaffrey, 2009). Moreover, unlike percentiles, mathematical operations could validly be applied to NCEs due to their scaling. See the Appendix for a percentile versus NCE comparison.

Ethics, Validity, and Reliability

There is no shortage of educational research from which to draw inferences or to develop one’s own study. However, one must examine research and purported conclusions with caution. It is important to scrutinize the methodologies employed to determine if the research met ethical
standards and produced results to a measure of reliability and validity. Throughout the study, the researcher took measures to address these three areas.

**Ethics**

Prior to designing the study, the researcher took a course in ethics as it specifically related to research involving human subjects. The researcher then submitted a study proposal to the Institutional Review Board, which determined the proposition included sufficient safeguarding to proceed. The *ex post facto* nature of the study alleviated many ethical concerns, as no researcher-subject interaction took place. The lack of both treatment manipulation and assignment also assured no inherent ethical dilemmas (Cohen, Manion, & Morrison, 2007). Coding student and teacher level data to remove any personal identifiers all but guaranteed there would be no breach of confidentiality, even in the unlikely event of the data file being obtained by an individual other than the researcher. Guidelines and requirements set forth in the Family Educational Rights and Privacy Act (FERPA) were followed at all times. Furthermore, even though individual student data were used during the study, the final research reporting only involved aggregate results. Additionally, while specific teacher variables were included in the analysis, the teacher effects were not the interest of the study. Rather, they were included to account for any variation due to teachers so that the actual independent variable of interest could be appropriately evaluated. Results could otherwise be confounded.

**Validity**

State departments of public instruction, including the NCDPI, use annual standardized state tests in particular subjects to provide a metric regarding the level of learning acquired by each student. Such assessments are criterion-referenced in that their primary intent is to provide information regarding to what extent a student has demonstrated mastery of specific content
standards. The Division of Accountability Services within the NCDPI creates these multiple choice tests, proceeding through a rigorous process of item development, review, tryouts, field tests, parallel item creation, pilot testing, and finally standards establishment (Creswell, 2012). While analyzing field test results, the NCDPI includes several other organizations to ascertain item appropriateness. Test developers create a mix of low, medium, or high levels of difficulty items, that when taken together, accurately reflect the state Standard Course of Study (SCOS). The SCOS communicates to teachers what students are to know and be able to do by the end of one year in the designated course. Only test items directly related to the SCOS appear on EOGs. Additionally, test developers publish test specifications documents, which inform teachers of the approximate percent of EOG items included for each standard. Furthermore, the NCDPI releases sample EOGs from which teachers can gain an even deeper sense of SCOS expectations as related to standardized test questions. Ultimately, multiple forms for each test are administered, equated statistically, and content-balanced. As such, the construct validity for the NC EOGs was deemed acceptable and essentially subsumed content validity as well (Cohen et al., 2007).

Educational research rarely employs truly randomized experiments, largely due to the involvement of human subjects, many of whom are children. While the majority of subjects comprising this study effectively received each treatment, it also was not an experimental design. Nevertheless, randomization is a particularly important aspect of sound statistical design. Yet even when random selection and/or assignment is not possible, one can mitigate the threat to validity by comprising groups of individuals as similar to one another as possible on the front end of the design (Creswell, 2012; Gay et al., 2009). In this study, the researcher matched individuals with themselves and selected a cohort of students who experienced stability in both teachers and administrators over the observed time period. Additionally, prior achievement was
used as a covariate during analysis to further control for initial differences, a critical component to *ex post facto* research (Spector, 1981).

By testing the statistical model with and without various potential fixed effects, random effects, and covariates, the researcher sought to enhance the study’s internal validity. While not possible to know of all variables or situations for which a control may be warranted, including analyses with those known led to more precise statements regarding possible relationships between and among variables (Gay et al., 2009).

**Reliability**

The same developers created the 7th and 8th grade EOGs, which went through identical rigorous procedures prior to being operational. The researcher mitigated the distorting impact of extreme observations by converting results to NCEs. For these two primary reasons, the EOG results were reliable indicators of student performance. Additionally, in that NC accountability models already incorporated EOG results, this study’s concern was not of whether the metrics were valid or reliable in and of themselves, but rather whether concern of their use was warranted due to a relationship between EOG results and factors unrelated to the teacher’s actions. The NCDPI incorporates multiple versions of each test and statistically analyzes results annually to produce stable results. While not pertinent to this study, student percentile values, converted here to NCE values, are always presented against the same norming year, thereby allowing 8th grade results one year to be reasonably compared to 8th grade results from a subsequent or prior year with the same norming base. The lack of subjectivity present in this research also contributed to the reliability of the results.

By definition, student performances on standardized tests are approximately normally distributed and have been so in reliable and consistent ways (Yates et al., 2002). As such, the
data for this study, in conjunction with analysis by sophisticated statistical software, were sufficiently reliable. Some anomalies in the data were expected. Utilizing sufficiently large samples for each subpopulation’s analysis and performing multiple statistical tests contributed positively to the reliability of the results.

Possible Threats to Reliability and Validity

Regression as an internal validity threat. The threat of regression cannot be controlled for as much as the researcher must be cognizant of the tendency for performers to regress to the mean over time. In general, this means when students perform above average one year, they would be predicted to score less high the next. The same would be expected for students who scored below average one year—they would be expected to perform better, in a relative sense, the next time (Gelman & Hill, 2007). Specifically in this study, students with above or below average change scores may in part be explained by this natural phenomenon—that they had opposite change scores the year prior. Notably, if the change in performance from 7th grade to 8th grade were essentially random, due to regression to the mean, one would expect a correlation in the proximity of $-\sqrt{0.71}$ (D. Spiegelhalter, personal communication, July 22, 2014). The researcher compared the actual study’s data correlation to the mean regressive expectation. As Allison (1990) described, given “…the almost universal phenomenon of regression toward the mean from pretest to posttest measurements…”, a model for change over time will typically show a negative correlation (p. 95). Practically, this means lower-level students should perform higher than in the past and high-level students should perform lower than in the past. Ostensibly then, should the low students perform lower than expected or the high level students perform even higher than expected, the impact of student-to-classroom assignment practices may be more significant than indicated by a $p$-value alone, or other variables may be interacting and impacting
the results. The use of a multilevel model (MLM) further mitigates the regression to the mean factor. Essentially, MLMs shrink parameter estimates and do so proportionally to the sample sizes from which they are derived. This conservative method of estimation effectively incorporates regressive tendencies and utilizes sample size to maintain palatable levels of bias (Bickel, 2007). The matched pairs nature of the data further mediated the threat to internal validity due to regression, as when initial similarities are controlled, results are less biased (Campbell & Stanley, 1963).

**Interaction with selection.** Another threat to this study’s internal validity stemmed from interaction with selection (Creswell, 2012). The researcher attempted to counteract this threat in two primary ways. First, the researcher included a matched pairs design in which student participants served as their own matched pair, or control, over time. Second, variables suspected of offering explanatory value to the variation in student achievement were sequentially analyzed for contribution. Those variables providing the most statistically significant contributions were included until the model lost power. Additionally, the researcher specifically used students’ prior testing history as a covariate for the study, a recommendation when research occurs with intact groups (Gay et al., 2009; Spector, 1981). Due to this study’s involvement of students in different classrooms with different teachers one year moving to different groupings with different teachers the next year, selection-maturation-interaction represented another threat to this study’s validity, particularly in a repeated measures, cross-classification design (Campbell & Stanley, 1963). While still present, the researcher mitigated this threat by including possible classroom and teacher effects in the analysis, including a latent factor for prior year teacher. For instance, each student’s two-year teacher sequence was included as an effect. Additionally, given the innate variability in individual annual achievement test results, the covariate representing student’s
entering ability was the average, when available, of their two year’s prior achievement test result. As such, the averaging limited the effect of a bad or good testing day and combined some portions of the grade six and seven teacher and peer effects as well.

**External validity threats.** The use of existing structures and data sets as opposed to active treatment application and data collection minimized threats to external validity. Adding an element of randomness to the selection of student data would enhance both generalizability and reliability. While students were mostly randomly allocated to their respective assignments via computer, some scheduling constraints limited the randomness of the outcome. Absent being able to truly randomly assign students to classrooms, having a larger sample of students, additional cohorts, or additional teachers would be an improvement to the study. While the quantitative nature of the study enhanced generalizability, the lack of randomness within the study and the use of a population within an urban magnet school both contributed to questionable generalizability (Cohen et al., 2007; Creswell, 2012). Given sound procedures, the study would be replicable and results could be applicable to past and future cohorts of students at this school, at a minimum, and could also extend to similar urban secondary schools.

**Analysis and Statistical Procedures**

**Dependent and Independent Variables**

The study’s primary independent variable of interest was student-to-classroom assignment practice—in other words, whether students were assigned to classes tracked by ability/performance or heterogeneously/mixed without regard to ability/performance. As such, this independent variable had two levels, tracked and mixed. A secondary independent variable of interest was level of student, taking on two values: advanced and non-advanced. Putting these two variables together within the school’s actualization of classroom assignment practices
between 7th and 8th grade ELA classes resulted in three “treatment groups,” advanced students who were tracked in 7th grade and mixed in 8th grade, non-advanced students who were tracked in 7th grade and mixed in 8th grade, and non-advanced students who were tracked for both 7th and 8th grade. The study hypothesized student-to-classroom assignment practices related to relative student achievement on state standardized tests. Therefore, the dependent variable of interest for the study was student relative achievement on North Carolina state English/Language Arts standardized tests.

**Initial Data Analysis**

After data collection, electronic spreadsheets were used to organize, de-identify, code, and clean the data in preparation for analysis. Due to the complexities of the study and the multitude of tests required, the researcher used statistical software to complete analyses, namely the Statistical Package for the Social Sciences (SPSS). Individual student achievement results for each of 7th and 8th grade ELA EOGs were converted to Normal Curve Equivalent units. To do so, percentile values were first converted to standardized Z-score values. NCEs were calculated according to the following formula:

\[ \text{NCE} = Z_{\text{percentile}} \times 21 + 50 \]  

(1)

where 21 and 50 represented the standard deviation and mean of NCEs, respectively.

NCEs, student demographic information, grouping trajectories, and teacher class period assignments were defined and coded in SPSS. How student achievement changed in 8th grade in NCE units was the data upon which primary analysis occurred. For each of the three subpopulations, advanced to mixed, non-advanced to mixed, and non-advanced staying tracked,
the researcher computed descriptive statistics and plotted achievement results to examine for patterns and shape. A preliminary analysis of variance (ANOVA) F test determined if the three groups’ mean differences were significantly different from one another. Analysis of Covariance (ANCOVA) tests examined the relationships among the three groups while controlling for suspected covariates. Additionally, a series of matched pairs t-tests examined student subgroup relative achievement in 8th grade based on 7th grade achievement.

**Modeling the Data**

While an ANOVA can alert the researcher to potentially significant mean differences, and ANCOVA can add control factors, more sophisticated statistical techniques offered the possibility of explaining and predicting those differences. One such option was the more complex multilevel model (MLM), also referred to as hierarchical models (HLM). MLMs are particularly useful in educational settings where, as in this case, the data of interest come from students nested within classrooms (Raudenbush, 2009; Raudenbush & Bryk, 2002; Singer & Willett, 2003). Such models were also known as mixed models in that both fixed and random effects were incorporated. The researcher built an MLM to allow for the inclusion of additional variables, such as student race, gender, prior performance, etc., and the modeling of multiple dependencies (Ballou, Sanders, & Wright, 2004). In this study, level one specifically modeled student outcomes while level two modeled classroom outcomes. Level one examined within group (WG) differences attributable to student level predictors while level two examined between group (BG) differences attributable to group/classroom level predictors. Using an MLM allowed for the explicit modeling of variance and covariance, splitting variation into WG and BG components. In addition to specific classrooms at level two, which accounted for possible
differential peer effects on outcomes, the predictor “type of class,” tracked or mixed, was examined for predictability along with other aggregated student characteristics.

Furthermore, in that specific teacher effects were not important in this study, yet teacher effects were potentially significant contributors to variation in student achievement, grade 7-8 teacher trajectories were included as student-level predictors. For example, students had either of teacher A or B in 7\textsuperscript{th} grade followed by either of teacher C or D in 8\textsuperscript{th} grade, creating four possible trajectories: AC, AD, BC, or BD. The researcher tested the model with various combinations of independent variables to determine which ones effectively contributed to the explanation of variance. The researcher included such variables in the final model and discarded others. Ultimately, the goal of the study was to model how the dependent variable, relative student achievement, was related to the primary independent variable, student-to-classroom assignment practice. MLMs offered the best way to accomplish this goal (Gay et al., 2009; Raudenbush & Bryk, 2002; Singer & Willett, 2003).

**Building a Multilevel Model**

Building a multilevel model involves a number of steps and various specifications. The majority of this study’s MLM discussion will occur in Chapter 4. However, some of the general procedures and concepts will first be presented here. The primary reason this study employed an MLM related to the nature of the data analyzed. Schools offer fundamental examples of hierarchically situated data. In this study, students produced the outcome data points of interest, and the students were nested in classrooms. The nesting feature drives the MLM, necessitating, in this case, two levels of models: one to model student level outcomes from student level predictors and another to model classroom level outcomes via classroom level predictors. Essentially, one can conceptualize an MLM as level two outcomes becoming the inputs for the
level one model, though in actuality everything happens simultaneously. Some researchers contended if a nesting structure were present, an MLM automatically should be used (Hoffman, 2015). Given a hierarchical situation, one way to quantify the need to employ an MLM was to partition the residual variance into components deriving from individuals and from the grouping structure. In this study, this amounted to determining how much of the variation in student outcomes was attributable to students individually, or within group (WG), and how much could be attributed to the classrooms in which they learned, or between group (BG). If a significant proportion of the variation, known as the Intraclass Correlation (ICC), were due to BG factors, MLMs provided the best method of analysis (Raudenbush & Bryk, 2002; Singer & Willett, 2003). To partition the variance, an empty means, random intercept model was used.

Empty or unconditional means, random intercept model:

\[
\text{Level 1: } Y_{ij} = \beta_{0j} + \varepsilon_{ij} \tag{2}
\]

where \( Y_{ij} \) represented the outcome for student \( i \) in class \( j \) and was determined by a combination of \( \beta_{0j} \), which represented the average outcome intercept, or average change score for class \( j \) and \( \varepsilon_{ij} \), the residual difference between the student \( i \) outcome and their class \( j \) average.

\[
\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j} \tag{3}
\]

where the intercept from Level 1, \( \beta_{0j} \), was modeled as the combination of \( \gamma_{00} \), or the
overall grand mean outcome across all groups/classrooms, and \( u_{0j} \), the difference between class \( j \)'s average outcome and the overall average outcome.

By substituting the Level 2 equation into the Level 1 equation, the MLM becomes:

\[
\text{Combined Model: } \gamma_{ij} = (\gamma_{00} + u_{0j}) + \epsilon_{ij} \tag{4}
\]

The WG variation of \( \epsilon_{ij} \)'s represented an individual student’s deviation from their class’ average and was measured as the residual variance, \( \sigma^2 \), assumed \( \sim N(0, \sigma^2) \). Similarly, the BG variation of \( u_{0j} \)'s represented an individual classroom’s deviation from the overall school average outcome and was measured as the intercept variance, \( \tau^2 \), also assumed \( \sim N(0, \tau^2) \). Additionally, the \( \epsilon_{ij} \)'s and \( u_{0j} \)'s were assumed to be independent. It follows that the Intraclass Correlation is:

\[
\text{ICC} = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{5}
\]

**Adding Predictors to the MLM**

After determining the extent to which the nesting or grouping structure, here classrooms, accounted for variation in student outcomes, Level 1 and Level 2 predictors were entered into the MLM to account for or explain portions of the remaining variation. This type of forward selection of predictors is a common practice with MLM specification. As with regression analyses, decisions regarding whether and how to center predictor variables must be addressed when employing an MLM. One of the main purposes in centering predictors is to establish a meaningful context for when the predictor value is zero. In the case of categorical predictors with more than two levels, a series of dummy codes were created. When variables are dummy coded,
each level of a predictor becomes its own variable, forcing all other levels to zero and using a one to indicate the level of interest. As such, there is no centering, per se. In the case of binary predictors, again there is not centering for Level 1, as one value is set to zero and the other to one. However, at Level 2, if such binary predictors are aggregated to each Level 2 group, the values should be grand mean centered (Hoffman, 2015). To grand mean center, the overall proportion of individuals with the characteristic coded as one at Level 1 is subtracted from each group’s average value for the same binary predictor, or the group proportion. With this type of centering, when both levels of predictors are included in the model, the Level 1 coefficient provides the within-group effect information while the Level 2 coefficient communicates the contextual effect of the predictor. A contextual effect describes what the extra benefit/detriment of being in that particular group is after controlling for the individual effect. On the other hand, group-mean centering offers Level 1 coefficients that provide within group effects and Level 2 coefficients that provide between group effects, and they are not dependent on one another for interpretation. Which kind of centering to use for the predictors is an empirical question. Again, while the specifics of the MLM will be included in Chapter 4, in general, after adding the predictor X, the Level 1 equation became:

\[ Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \epsilon_{ij} \]  

(6)

Note the only difference is the inclusion of \( \beta_{1j} X_{ij} \), where \( X_{ij} \) represented the level 1 predictor value for student \( i \) from classroom \( j \) and \( \beta_{1j} \) signified the slope or rate of change for predictor 1 in classroom \( j \).
The level 2 equations model the slopes:

\[
\begin{align*}
\text{Level 2 slope equations:} & \quad \beta_{0j} = \gamma_{00} + u_{0j}, \text{ as before, and} \\
& \quad \beta_{1j} = \gamma_{10} \\
\end{align*}
\]

which would indicate a fixed effect of \(\gamma_{10}\) for all classrooms; alternatively, a level 2 predictor could be included to model \(\beta_{0j}\), as:

\[
\beta_{0j} = \gamma_{00} + \gamma_{01}X_{\cdot j} + u_{0j}
\]

where \(X_{\cdot j}\) represented the average class \(j\) value for predictor 1, and \(\gamma_{01}\) indicated how the value of \(X_{\cdot j}\) affected class \(j\)’s overall outcome.

Furthermore, the \(\beta_{1j}\) could vary such that yet another model would be needed:

\[
\beta_{1j} = \gamma_{10} + \gamma_{11}X_{\cdot j} + u_{1j}
\]

Note the only difference here is in the subscripts, which are now \(10\) or \(1j\), indicating the random slope was now the outcome as opposed to the random intercept.

By substituting equations (7) and (8) into (5), we end up with an MLM including a fixed intercept, fixed level 1 predictor, fixed level 2 predictor, a cross-level interaction, and both a random slope and random intercept.
\[ Y_{ij} = \gamma_{00} + \gamma_{01}X \cdot j + u_{0j} + (\gamma_{10} + \gamma_{11}X \cdot j + u_{1j}) X_{ij} + \varepsilon_{ij} \quad \text{simplifies to} \quad (10) \]

\[ Y_{ij} = \gamma_{00} + \gamma_{01}X \cdot j + \gamma_{10} X_{ij} + \gamma_{11} X \cdot j X_{ij} + u_{1j} X_{ij} + u_{0j} + \varepsilon_{ij} \quad (11) \]

It is evident, then, how quickly degrees of freedom can be encumbered via parameter estimation, as including one predictor at both levels and allowing variation for one intercept and one slope required the estimation of at least six parameters. As such, it is essential to run various model specifications to capture the most significant effects prior to either saturating the model or losing enough power to reach model convergence.

**Significance Level**

The researcher followed convention, naming a significance level, \( \alpha \), of 0.05 for the study. Essentially, this meant if results were less than 5% likely to occur by pure chance, the null hypothesis being tested would be rejected. Such rejections, by definition, would be incorrect 5% of the time, also known as the probability of making a Type 1 error (Yates et al., 2002). In other words, researchers effectively incorrectly reject null hypotheses (\( \alpha \times 100 \)%) of the time.

Nonetheless, throughout the study, the difference between theoretically statistically significant versus practically significant results must also be considered. In other words, distinguishing between a result with a \( p \)-value of 0.046 and one of 0.054 can be seen as arbitrary. Therefore, results will be interpreted in context and evaluated multiple ways for their contribution to the overall model. Additionally, as predictors are added to the MLM, particularly if they are collinear with an already included predictor, effects and therefore significance levels can shift.
Rationale for Statistical Procedures

As indicated by Creswell (2012), when a researcher utilized intact groups, as was the case in this study, the design can only be quasi-experimental. However, this study was more of an *ex post facto* correlational design. While there were different treatments, tracked versus mixed level ability classes, the researcher did not manipulate those treatments. ANCOVA F tests were appropriate when comparing more than two populations of interest, taking into account both the means and variances from the sample data, while also including a likely significant covariate (Wright, 2006). Such a test illuminated whether any differences existed between any of the populations of interest. However, to capture and model the complexity inherent in the study, a multilevel model was needed (Cohen et al., 2007). MLMs offered more flexibility than ANCOVA and matched pairs t-tests alone, accounted for the hierarchical nature of the sample, and could actually model the variance and covariance components of the model (Quené & Van den Bergh, 2004). While possible to include all variables simultaneously in a one-level multiple regression model, doing so often under-estimated variance, leading to biased results. Additionally, and critical to study conclusions, MLMs treated data sources as random samples from populations rather than simply as fixed data of interest. The implications of this difference were far reaching: by including the modeling of the variance and not solely the fixed effects, one could begin to answer the question of why group A tended to be higher than group B, and conclusions could apply to the greater population of groups from which they were obtained. In this study, the goal was not to simply identify differences but to explain why identified differences occurred such that one could predict differences based on similar characteristics moving forward.
Limitations and Significance

As with any study, there were a number of limitations affecting the study itself and its subsequent applicability. Nonetheless, the significance of this research and its contribution to filling a gap in the literature remain.

Limitations

A primary limitation to this study was its consideration of student-to-classroom assignment practices’ relationship only with relative student achievement. Many other dependent variables of interest existed, such as the demographic composition of various levels of courses or actual student performance, classroom grades, student classroom engagement level, student and teacher perceptions, etc. Education sets out many and sometimes competing goals for its students and teachers. Therefore, looking at other quantitative variables would add a broader and more authentic picture of what the relationships of student-to-classroom assignment practices with other variables, including relative student achievement, may be. Additionally, interviews of students, parents, teachers and administrators could further contextualize and offer insight into the less obviously measureable outcomes realized in various grouping methods. Including this mixed methods analysis of feelings and perceptions would add dimension to the study.

Moreover, while many states have moved toward using relative student achievement data to essentially rate schools and teachers, not all have done so, making the implications of the findings possibly less relevant to those particular states.

Given the dynamic nature of the teaching and learning exchange between and among teachers and students, the generalizability of any educational study is questionable. While quantitative studies, due to their larger sample sizes and more objective variables, have more transferable results, one could argue this study would only apply to this specific school or to
similar urban secondary schools, thereby limiting its possible influence. Another limitation was the lack of experimental design possibility. While not practical, if the same students had been assigned to both treatments from the same teachers, the study’s results would be more compelling. Completing the analysis one time and with a single cohort of students was another limitation of the study.

**Significance**

For two primary reasons, this study’s results were significant. First, if student-to-classroom assignment practices have a relationship with relative student achievement, organizational adjustments could be made to optimize student learning. This outcome, of course, assumes standardized test metrics to be a valid indicator of student learning. Relatedly, given the segregative effect of tracking students by ability and the associated achievement gap, if student learning can be positively affected by altering organizational practices, more equitable outcomes may result. The impact of greater educational equity would reach far beyond the school walls.

Second, if student-to-classroom assignment practices have a relationship with relative student achievement and those results were used to assess teacher and school effectiveness, modifications would have to be made to teacher evaluation systems to reflect the fact that at least some of the results arose from structural factors not under the teacher’s control.

**Chapter Summary**

This chapter included an overview of the methodology for the study. Rationale for engaging in a quantitative design along with intended sample sizes and analytical procedures were presented. Significant precautions were taken to maintain the anonymity of all data sources, yielding no known risks inherent within the study design. The analytical results of the study are presented in the next chapter.
CHAPTER FOUR: RESULTS

Introduction

This chapter presents the results of the analysis regarding possible relationships between how students were assigned to classrooms for English Language Arts (ELA) and their relative student achievement on standardized state assessments. First, descriptive and summary statistics about the students and the classes are offered. Second, analytical results from several methodological approaches are presented.

Descriptive Statistics

The primary response variable of interest in this study was students’ relative performance on the 8th grade ELA End of Grade (EOG) state standardized test. While the 8th grade result was the dependent variable, only students who attended the school for both 7th and 8th grades were included in the analysis. This decision resulted from the primary explanatory or independent variable in the study, student to classroom assignment method. In 7th grade all students at the school were tracked by ability for ELA instruction. When they moved to 8th grade, all of the advanced students and 70% of the non-advanced students were assigned to mixed level ELA classes, and the remaining non-advanced students remained tracked into one of two lower level ELA classes. There was no difference between the non-advanced students assigned to mixed classes and those assigned to the tracked classes with the exception of the students with disabilities who required inclusion services. They were automatically assigned to one of the two tracked 8th grade classes, both of which offered inclusion services wherein an additional teacher
specifically supported their ELA classroom experience. Consequently, relative student achievement, that is, how students performed on the 8th grade EOG as compared to how they performed on the 7th grade EOG, coincided with a change from tracked to mixed level instruction or the maintenance of low level tracking. The overall cohort demographics are presented in Table 1. The aspect of the study necessitating the employment of a multilevel model (MLM) analysis was that students were nested in classrooms, possibly conflating the influence of the treatment, tracked or mixed level, on EOG performance. Therefore, descriptive statistics for the eight 8th grade ELA classrooms are also presented in Table 2.

Table 1

*Overall Cohort Descriptive Statistics*

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>131</td>
<td>68.6</td>
</tr>
<tr>
<td>Male</td>
<td>60</td>
<td>31.4</td>
</tr>
<tr>
<td>African American</td>
<td>81</td>
<td>42.4</td>
</tr>
<tr>
<td>White</td>
<td>50</td>
<td>26.2</td>
</tr>
<tr>
<td>Latino</td>
<td>45</td>
<td>23.6</td>
</tr>
<tr>
<td>Other</td>
<td>15</td>
<td>7.9</td>
</tr>
<tr>
<td>Academically Intellectually Gifted</td>
<td>74</td>
<td>38.7</td>
</tr>
<tr>
<td>Student With Disability</td>
<td>15</td>
<td>7.9</td>
</tr>
<tr>
<td>English Language Learner</td>
<td>13</td>
<td>6.8</td>
</tr>
</tbody>
</table>
Table 2

*Descriptive Statistics by 8th Grade Class Period*

<table>
<thead>
<tr>
<th>Class Period Gr 8</th>
<th>Tracked</th>
<th>Tracked</th>
<th>Mixed</th>
<th>Mixed</th>
<th>Mixed</th>
<th>Mixed</th>
<th>Mixed</th>
<th>Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion Male</td>
<td>.25&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.40&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.19&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.22&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.30&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.42&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.43&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.33&lt;sub&gt;a&lt;/sub&gt;</td>
</tr>
<tr>
<td>Proportion Minority</td>
<td>.80&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.73&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.70&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.70&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.79&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.79&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.70&lt;sub&gt;a&lt;/sub&gt;</td>
<td></td>
</tr>
<tr>
<td>Proportion AIG</td>
<td>.00&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.20&lt;sub&gt;a,b&lt;/sub&gt;</td>
<td>.41&lt;sub&gt;a,b&lt;/sub&gt;</td>
<td>.30&lt;sub&gt;a,b&lt;/sub&gt;</td>
<td>.57&lt;sub&gt;b&lt;/sub&gt;</td>
<td>.54&lt;sub&gt;b,c&lt;/sub&gt;</td>
<td>.46&lt;sub&gt;b,d&lt;/sub&gt;</td>
<td>.48&lt;sub&gt;b,e&lt;/sub&gt;</td>
</tr>
<tr>
<td>Proportion 7th Grade Teacher A</td>
<td>.75&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.60&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.41&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.52&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.35&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.29&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.43&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.63&lt;sub&gt;a&lt;/sub&gt;</td>
</tr>
<tr>
<td>Proportion Non-Advanced</td>
<td>1.00&lt;sub&gt;a&lt;/sub&gt;</td>
<td>1.00&lt;sub&gt;a,b&lt;/sub&gt;</td>
<td>.56&lt;sub&gt;b,c&lt;/sub&gt;</td>
<td>.63&lt;sub&gt;a,b,c&lt;/sub&gt;</td>
<td>.39&lt;sub&gt;c&lt;/sub&gt;</td>
<td>.50&lt;sub&gt;c,d&lt;/sub&gt;</td>
<td>.50&lt;sub&gt;c,e&lt;/sub&gt;</td>
<td>.56&lt;sub&gt;b,c,f&lt;/sub&gt;</td>
</tr>
<tr>
<td>Proportion with Support Class</td>
<td>.40&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.33&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.26&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.15&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.09&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.17&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.18&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.07&lt;sub&gt;a&lt;/sub&gt;</td>
</tr>
<tr>
<td>Proportion Below Average Starting 8th</td>
<td>.75&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.50&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.37&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.37&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.30&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.33&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.36&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.31&lt;sub&gt;a&lt;/sub&gt;</td>
</tr>
<tr>
<td>Average 2013 Level, 1-4</td>
<td>2.20&lt;sub&gt;a&lt;/sub&gt;</td>
<td>2.27&lt;sub&gt;a,b&lt;/sub&gt;</td>
<td>2.78&lt;sub&gt;a,b&lt;/sub&gt;</td>
<td>2.62&lt;sub&gt;a,b&lt;/sub&gt;</td>
<td>2.91&lt;sub&gt;a,b&lt;/sub&gt;</td>
<td>3.04&lt;sub&gt;b&lt;/sub&gt;</td>
<td>2.75&lt;sub&gt;a,b&lt;/sub&gt;</td>
<td>2.63&lt;sub&gt;a,b&lt;/sub&gt;</td>
</tr>
<tr>
<td>Proportion 8th Grade Teacher C</td>
<td>1.00&lt;sup&gt;1&lt;/sup&gt;</td>
<td>.00&lt;sup&gt;1&lt;/sup&gt;</td>
<td>.00&lt;sup&gt;1&lt;/sup&gt;</td>
<td>1.00&lt;sup&gt;1&lt;/sup&gt;</td>
<td>.00&lt;sup&gt;1&lt;/sup&gt;</td>
<td>.00&lt;sup&gt;1&lt;/sup&gt;</td>
<td>1.00&lt;sup&gt;1&lt;/sup&gt;</td>
<td>1.00&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

1. This category is not used in comparisons because there are no other valid categories to compare.
2. Tests are adjusted for all pairwise comparisons within a row using the Bonferroni correction.

As students moved from 7th to 8th grade, they were part of one of three possible grouping trajectories: advanced to mixed, non-advanced to mixed, and non-advanced stay tracked. The demographics for the three grouping trajectories are presented in Table 3.

Table 3

*Descriptive Statistics by Grouping Sequence Trajectory*

<table>
<thead>
<tr>
<th>Grouping Sequence</th>
<th>Non-advanced stay tracked</th>
<th>Non-advanced to mixed</th>
<th>Advanced to mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion Male</td>
<td>.31</td>
<td>.33</td>
<td>.30</td>
</tr>
<tr>
<td>Proportion Minority</td>
<td>.77</td>
<td>.94</td>
<td>.50</td>
</tr>
<tr>
<td>Proportion AIG</td>
<td>.09</td>
<td>.11</td>
<td>.84</td>
</tr>
<tr>
<td>Proportion with Support Class</td>
<td>.37</td>
<td>.29</td>
<td>.00</td>
</tr>
<tr>
<td>Proportion Below Average</td>
<td>.65</td>
<td>.61</td>
<td>.04</td>
</tr>
<tr>
<td>Starting 8th</td>
<td>.23</td>
<td>.32</td>
<td>.30</td>
</tr>
<tr>
<td>Average 2013 Level, 1-4</td>
<td>2.23</td>
<td>2.32</td>
<td>3.30</td>
</tr>
</tbody>
</table>

107
Given the concern regarding ability grouping mirroring segregation, only the racial demographics by class period are presented by class period, for 7th grade reflecting three advanced sections and five non-advanced sections, and for 8th grade, reflecting six mixed ability sections and two non-advanced sections of ELA. This comparison of proportion minority by class period is presented in Table 4.

Table 4

*Proportion Minority by Class Period, Grades 7 and 8*

<table>
<thead>
<tr>
<th>Proportion Minority</th>
<th>Class Period Gr 7</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>.79&lt;sub&gt;a,c,d&lt;/sub&gt;</td>
<td>.60&lt;sub&gt;a,b,c,f&lt;/sub&gt;</td>
<td>.38&lt;sub&gt;b&lt;/sub&gt;</td>
<td>1.00&lt;sub&gt;c&lt;/sub&gt;</td>
<td>.90&lt;sub&gt;a,c&lt;/sub&gt;</td>
<td>.46&lt;sub&gt;b,d&lt;/sub&gt;</td>
<td>.91&lt;sub&gt;c,e&lt;/sub&gt;</td>
<td>.90&lt;sub&gt;c,f&lt;/sub&gt;</td>
<td></td>
</tr>
<tr>
<td>Class Period Gr 8</td>
<td>Tracked</td>
<td>Tracked</td>
<td>Mixed</td>
<td>Mixed</td>
<td>Mixed</td>
<td>Mixed</td>
<td>Mixed</td>
<td>Mixed</td>
</tr>
<tr>
<td>Proportion Minority</td>
<td>.80&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.73&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.70&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.70&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.70&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.79&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.79&lt;sub&gt;a&lt;/sub&gt;</td>
<td>.70&lt;sub&gt;a&lt;/sub&gt;</td>
</tr>
</tbody>
</table>

1. Tests are adjusted for all pairwise comparisons within a row using the Bonferroni correction.

To confirm the apparent dependency between 7th grade track assignment and race, a Chi-square test for independence was utilized. The results indicated the distribution of race was not independent of 7th grade class period. Specifically, $\chi^2 (7, N = 186) = 47.57, p < 0.000$, meaning if race were not related to class period, the observed results essentially would never occur. In contrast, there was no evidence of racial class period dependency in 8th grade, yielding $\chi^2 (7, N = 191) = 1.79, p = 0.970$. Essentially, these results indicate a segregative impact of tracking on classroom composition.

**Performance Metrics**

In addition to investigating the various demographics and other indicator statistics by class period, race, and grouping sequence, various performance metrics were explored. The three main categories of interest were students’ average EOG scores from grades six and seven when all students were tracked, their eighth grade EOG results when most students were mixed by
ability for instruction, and the difference between their eighth grade results and their sixth-seventh grade averages, a version of a change score. As a reminder, Normal Curve Equivalents (NCEs) are the commonly scaled metric representing EOG assessment results. The scale for NCEs runs from one to 99. In addition to examining simple mean scores, medians are reported, as they are resistant to the effect of outliers, along with indicators of dispersion, including minimum, maximum, range, and standard deviation. These performance metrics are presented for each of White, African American, and Latino students in Table 5.
Table 5

*Performance Metrics by Race*

<table>
<thead>
<tr>
<th>Race</th>
<th>White</th>
<th>African American</th>
<th>Latino</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average 6/7 NCE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>50</td>
<td>81</td>
<td>45</td>
</tr>
<tr>
<td>Mean</td>
<td>70.08</td>
<td>49.70</td>
<td>45.20</td>
</tr>
<tr>
<td>Maximum</td>
<td>99.00</td>
<td>84.50</td>
<td>82.00</td>
</tr>
<tr>
<td>Median</td>
<td>72.00</td>
<td>50.00</td>
<td>46.50</td>
</tr>
<tr>
<td>Minimum</td>
<td>19.50</td>
<td>14.50</td>
<td>9.50</td>
</tr>
<tr>
<td>Range</td>
<td>79.50</td>
<td>70.00</td>
<td>72.50</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>16.79</td>
<td>15.57</td>
<td>16.65</td>
</tr>
<tr>
<td><strong>Grade 8 NCE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>72.54</td>
<td>52.84</td>
<td>48.97</td>
</tr>
<tr>
<td>Maximum</td>
<td>99.0</td>
<td>98.0</td>
<td>84.0</td>
</tr>
<tr>
<td>Median</td>
<td>76.0</td>
<td>53.0</td>
<td>50.5</td>
</tr>
<tr>
<td>Minimum</td>
<td>32.0</td>
<td>5.0</td>
<td>9.0</td>
</tr>
<tr>
<td>Range</td>
<td>67.0</td>
<td>93.0</td>
<td>75.0</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>16.5</td>
<td>16.9</td>
<td>15.6</td>
</tr>
<tr>
<td><strong>8 – 6/7avg NCE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>2.16</td>
<td>3.14</td>
<td>4.90</td>
</tr>
<tr>
<td>Maximum</td>
<td>24.00</td>
<td>27.00</td>
<td>25.00</td>
</tr>
<tr>
<td>Median</td>
<td>1.50</td>
<td>4.00</td>
<td>3.50</td>
</tr>
<tr>
<td>Minimum</td>
<td>-20.50</td>
<td>-22.50</td>
<td>-12.00</td>
</tr>
<tr>
<td>Range</td>
<td>44.50</td>
<td>49.50</td>
<td>37.00</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>10.56</td>
<td>11.17</td>
<td>8.80</td>
</tr>
</tbody>
</table>

As seen in Table 5, each racial subgroup had mean NCE increases between 7th and 8th grades. To assess whether these increases were statistically significant or plausibly due to chance, matched pairs t-tests were used. Based on 2-tailed tests, the increase for White students was $t(48, N = 49) = 1.43, p = 0.158$, for African American students, $t(81, N = 82) = 2.23, p = 0.028$, and for Latino students, $t(42, N = 43) = 3.65, p < 0.000$. Both African American and Latino students had statistically significant 8th grade NCE increases in overall average performance, while White students also gained, theirs was not statistically significant.
The relationship between how students were assigned to classrooms for instruction and their relative achievement on standardized tests was the primary area under investigation in this study. Therefore, similar preliminary exploration of student performance metrics by grouping trajectory, or how they were assigned to classes, was warranted. Analogously, Table 6 presents the same metrics as Table 5, this time for each of the three grouping trajectories.

Table 6

**Performance Metrics by Grouping Sequence**

<table>
<thead>
<tr>
<th>Grouping Sequence</th>
<th>Nadv 7 Tracked</th>
<th>Nadv 7 Mixed</th>
<th>Adv 7 Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>35 a</td>
<td>82 a</td>
<td>74 b</td>
</tr>
<tr>
<td>Average 6/7 NCE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>44.62 a</td>
<td>43.91 a</td>
<td>71.95 b</td>
</tr>
<tr>
<td>Maximum</td>
<td>86.00 a</td>
<td>67.50 a</td>
<td>99.00 b</td>
</tr>
<tr>
<td>Median</td>
<td>46.50 a</td>
<td>44.75 a</td>
<td>73.00 b</td>
</tr>
<tr>
<td>Minimum</td>
<td>9.50 a</td>
<td>10.50 a</td>
<td>45.00 b</td>
</tr>
<tr>
<td>Range</td>
<td>76.50 a</td>
<td>57.00 a</td>
<td>54.00 b</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>15.93 a</td>
<td>12.02 a</td>
<td>12.16 b</td>
</tr>
<tr>
<td>Grade 8 NCE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>46.04 a</td>
<td>49.41 a</td>
<td>73.93 b</td>
</tr>
<tr>
<td>Maximum</td>
<td>90.00 a</td>
<td>76.00 a</td>
<td>99.00 b</td>
</tr>
<tr>
<td>Median</td>
<td>50.00 a</td>
<td>48.00 a</td>
<td>76.00 b</td>
</tr>
<tr>
<td>Minimum</td>
<td>5.00 a</td>
<td>17.00 a</td>
<td>44.00 b</td>
</tr>
<tr>
<td>Range</td>
<td>85.00 a</td>
<td>59.00 a</td>
<td>55.00 b</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>17.83 a</td>
<td>12.80 a</td>
<td>13.30 b</td>
</tr>
<tr>
<td>8 – 6/7avg NCE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1.96 a</td>
<td>5.51 a</td>
<td>1.81 a</td>
</tr>
<tr>
<td>Maximum</td>
<td>25.00 a</td>
<td>30.00 a</td>
<td>24.00 a</td>
</tr>
<tr>
<td>Median</td>
<td>1.25 a</td>
<td>6.00 a</td>
<td>1.00 a</td>
</tr>
<tr>
<td>Minimum</td>
<td>-22.50 a</td>
<td>-18.00 a</td>
<td>-21.50 a</td>
</tr>
<tr>
<td>Range</td>
<td>47.50 a</td>
<td>48.00 a</td>
<td>45.50 a</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>10.78 a</td>
<td>10.30 a</td>
<td>10.24 a</td>
</tr>
</tbody>
</table>

1. Tests are adjusted for all pairwise comparisons within a row using the Bonferroni correction.
Again, all three subgroups had mean NCE increases between 7th and 8th grades, though the increase for the students in the non-advanced to mixed trajectory appears to be significantly higher than the other two average NCE increases. These differences can be seen graphically in Figure 3. The similar changes for advanced to mixed and non-advanced stay tracked students show up as the pair of lines with more similar slopes, while the substantial increase for non-advanced to mixed students is evidenced by the highly positive slope.

![Graph showing NCE change scores by grouping sequence](image)

*Figure 3. Visual representation of 8-6/7avg NCE change scores by grouping sequence.*

A unique element to this study was that through 7th and 8th grades, 81.6% of students experienced both tracking and mixing by ability ELA classroom assignment treatments. All students were in separate high or low level ELA classes in 7th grade. All advanced students and 70% of non-advanced students moved into mixed ability ELA classrooms in 8th grade. Therefore, students in these two grouping trajectories, advanced and non-advanced from tracked to mixed, could serve as their own matched pair in more of an experimental design sense. Their differences in performance, on an individual level, could be attributed to time-varying characteristics such as
how they were assigned to classes, as opposed to being attributable to any time-invariant characteristics, such as race, family background, sex, etc.

Matched pair 2-tailed tests for the advanced students moving to a mixed ability classroom indicated their average NCE increase of 1.81 was not statistically significant, $t(71, N = 72) = 1.50, p = 0.1374$. On the other hand, the average NCE gain of 5.51 for non-advanced students moving from tracked to mixed ability classes was highly statistically significant, $t(82, N = 83) = 4.70, p < 0.000$. The final group, the non-advanced students who stayed tracked for both years, can still legitimately comprise a matched pairs $t$-test, though this time there is no treatment change. Rather, it is of interest if staying tracked yielded a statistically significant change in average NCE score. For this group of students, the only time-varying characteristics would be classroom peers and teacher. Their gain of 1.96 was not statistically significant, $t(34, N = 35) = 1.06, p = 0.2984$. Overall, for this cohort of students, two of three time-varying elements were of interest to this study: classroom or peer effects and grouping structure. These two are related in that beyond random substantive changes in their peers occurred between 7th and 8th grade largely due to their simultaneous change from tracked to mixed level classrooms. The third time-varying element of classroom teacher had to be controlled for during the analysis. Absorbed in this matched pairs analysis is any possible influence of classroom peers. Therefore, even though the results point to a confirmation of the hypothesis that non-advanced students have better relative attainment when assigned to mixed level classrooms, it is necessary to further analyze the data by way of a multilevel model (MLM).

The nesting of students in classrooms results in homophily, or the tendency for students in a particular class to be more similar to one another than to students in another class. Multilevel models (MLMs) provide a way to account for such dependencies, allowing for the determination
of the amount of variation in achievement attributable to the group, in this case classroom, versus that attributable to the individual student. From a purely descriptive perspective, Table 7 presents information about each of the eight eighth grade class periods. At a cursory glance, there appear to be differences among the classroom outcomes. These discrepancies will be examined with the goal of explanation of variation later in this chapter.
Table 7

*Descriptive Performance Statistics by 8th Grade Class Period*

<table>
<thead>
<tr>
<th>Class Period Gr 8</th>
<th>Tracked1</th>
<th>Tracked2</th>
<th>Mixed1</th>
<th>Mixed2</th>
<th>Mixed3</th>
<th>Mixed4</th>
<th>Mixed5</th>
<th>Mixed6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>20\textsubscript{a}</td>
<td>15\textsubscript{a}</td>
<td>27\textsubscript{a}</td>
<td>27\textsubscript{a}</td>
<td>23\textsubscript{a}</td>
<td>24\textsubscript{a}</td>
<td>28\textsubscript{a}</td>
<td>27\textsubscript{a}</td>
</tr>
<tr>
<td>Average 6/7 NCE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>43.53\textsubscript{a}</td>
<td>46.18\textsubscript{a}</td>
<td>56.80\textsubscript{a}</td>
<td>53.65\textsubscript{a}</td>
<td>61.09\textsubscript{a}</td>
<td>58.04\textsubscript{a}</td>
<td>56.34\textsubscript{a}</td>
<td>57.52\textsubscript{a}</td>
</tr>
<tr>
<td>Mean</td>
<td>43.2\textsubscript{a}</td>
<td>49.9\textsubscript{b,a}</td>
<td>59.8\textsubscript{b}</td>
<td>57.2\textsubscript{a,b}</td>
<td>66.4\textsubscript{b,c}</td>
<td>65.8\textsubscript{b,d}</td>
<td>60.9\textsubscript{b,e}</td>
<td>56.9\textsubscript{a,b}</td>
</tr>
<tr>
<td>Maximum</td>
<td>76.00\textsubscript{a}</td>
<td>86.00\textsubscript{a}</td>
<td>90.50\textsubscript{a}</td>
<td>90.50\textsubscript{a}</td>
<td>89.00\textsubscript{a}</td>
<td>92.50\textsubscript{a}</td>
<td>83.00\textsubscript{a}</td>
<td>99.00\textsubscript{a}</td>
</tr>
<tr>
<td>Maximum</td>
<td>66.0\textsubscript{a}</td>
<td>90.0\textsubscript{a,b}</td>
<td>92.0\textsubscript{b}</td>
<td>87.0\textsubscript{a,b}</td>
<td>99.0\textsubscript{b,c}</td>
<td>98.0\textsubscript{b,d}</td>
<td>99.0\textsubscript{b,e}</td>
<td>99.0\textsubscript{a,b}</td>
</tr>
<tr>
<td>Median</td>
<td>44.00\textsubscript{a}</td>
<td>50.50\textsubscript{a}</td>
<td>55.50\textsubscript{a}</td>
<td>55.50\textsubscript{a}</td>
<td>60.50\textsubscript{a}</td>
<td>55.25\textsubscript{a}</td>
<td>54.25\textsubscript{a}</td>
<td>54.75\textsubscript{a}</td>
</tr>
<tr>
<td>Median</td>
<td>45.0\textsubscript{a}</td>
<td>53.0\textsubscript{a,b}</td>
<td>58.0\textsubscript{b}</td>
<td>59.0\textsubscript{a,b}</td>
<td>66.0\textsubscript{b,c}</td>
<td>66.0\textsubscript{b,d}</td>
<td>62.5\textsubscript{b,e}</td>
<td>53.0\textsubscript{a,b}</td>
</tr>
<tr>
<td>Minimum</td>
<td>19.50\textsubscript{a}</td>
<td>9.50\textsubscript{a}</td>
<td>24.00\textsubscript{a}</td>
<td>10.50\textsubscript{a}</td>
<td>39.50\textsubscript{a}</td>
<td>28.50\textsubscript{a}</td>
<td>23.00\textsubscript{a}</td>
<td>25.00\textsubscript{a}</td>
</tr>
<tr>
<td>Minimum</td>
<td>5.0\textsubscript{a}</td>
<td>9.0\textsubscript{a,b}</td>
<td>33.0\textsubscript{b}</td>
<td>17.0\textsubscript{a,b}</td>
<td>40.0\textsubscript{b,c}</td>
<td>32.0\textsubscript{b,d}</td>
<td>26.0\textsubscript{b,e}</td>
<td>30.0\textsubscript{a,b}</td>
</tr>
<tr>
<td>Range</td>
<td>56.50\textsubscript{a}</td>
<td>76.50\textsubscript{a}</td>
<td>66.50\textsubscript{a}</td>
<td>80.00\textsubscript{a}</td>
<td>49.50\textsubscript{a}</td>
<td>64.00\textsubscript{a}</td>
<td>60.00\textsubscript{a}</td>
<td>74.00\textsubscript{a}</td>
</tr>
<tr>
<td>Range</td>
<td>61.0\textsubscript{a}</td>
<td>81.0\textsubscript{a,b}</td>
<td>59.0\textsubscript{b}</td>
<td>70.0\textsubscript{a,b}</td>
<td>59.0\textsubscript{b,c}</td>
<td>66.0\textsubscript{b,d}</td>
<td>73.0\textsubscript{b,e}</td>
<td>69.0\textsubscript{a,b}</td>
</tr>
<tr>
<td>StdDev</td>
<td>13.82\textsubscript{a}</td>
<td>18.98\textsubscript{a}</td>
<td>20.05\textsubscript{a}</td>
<td>19.52\textsubscript{a}</td>
<td>15.41\textsubscript{a}</td>
<td>19.83\textsubscript{a}</td>
<td>17.92\textsubscript{a}</td>
<td>18.65\textsubscript{a}</td>
</tr>
<tr>
<td>StdDev</td>
<td>16.1\textsubscript{a}</td>
<td>19.7\textsubscript{a,b}</td>
<td>18.2\textsubscript{b}</td>
<td>19.7\textsubscript{a,b}</td>
<td>14.1\textsubscript{b,c}</td>
<td>18.0\textsubscript{b,d}</td>
<td>18.6\textsubscript{b,e}</td>
<td>17.4\textsubscript{a,b}</td>
</tr>
</tbody>
</table>

1. Tests are adjusted for all pairwise comparisons within a row using the Bonferroni correction.

As the primary interest of the study was to determine if there was a relationship between student assignment to ELA classes and relative student achievement, another appropriate initial
test was an Analysis of Variance (ANOVA) to see if the mean outcomes across the three grouping trajectories differed from each other in a statistically significant way. Running an ANOVA produces the same results as running a regression analysis. The results are presented in Table 8. While the results are not theoretically statistically significant at a 0.05 level, with a p-value of 0.059 they are practically significant, just 0.009 above the set alpha level. Additional exploration is warranted. A related Analysis of Covariance (ANCOVA) allows for a test of differences in means among groups while also conditioning for a potential covariate. Based on the literature, the most likely influential covariate is the effect of the classroom teacher. As such, the same ANOVA was run with eighth grade teacher included as a covariate, thereby running an ANCOVA. The results are presented in Table 9.

Table 8

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>2</td>
<td>307.810</td>
<td>2.866</td>
<td>.059</td>
</tr>
<tr>
<td>Within Groups</td>
<td>186</td>
<td>107.406</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>5</td>
<td>266.105</td>
<td>2.528</td>
<td>.031</td>
</tr>
<tr>
<td>Intercept</td>
<td>1</td>
<td>2016.068</td>
<td>19.155</td>
<td>.000</td>
</tr>
<tr>
<td>GroupingSequence</td>
<td>2</td>
<td>263.785</td>
<td>2.506</td>
<td>.084</td>
</tr>
<tr>
<td>Teacher138</td>
<td>1</td>
<td>618.020</td>
<td>5.872</td>
<td>.016</td>
</tr>
<tr>
<td>GroupingSequence * Teacher138</td>
<td>2</td>
<td>53.065</td>
<td>.504</td>
<td>.605</td>
</tr>
</tbody>
</table>
The corrected model $p$-value of 0.031 indicates a statistically significant difference across grouping trajectories in the relative student achievement, that is, the difference in NCEs between students’ 8th grade EOG results and their 6/7 average EOG scores. Additionally, the classroom teacher plays a significant role in the outcome. The flexibility afforded by a multilevel model to predict and condition for outcomes at both the classroom and student level will provide the most reliable analysis.

**Analysis via Multilevel Models**

Whenever data come from inherently hierarchical situations, such as students nested within classrooms, the most appropriate way in which to model and analyze the data is to use a multilevel model (MLM), also known as a mixed model, hierarchical linear model (HLM), and random coefficient model (Hoffman, 2015; Raudenbush & Bryk, 2002). The nesting feature within the data creates dependencies that would otherwise violate the assumptions required for standard regression. MLMs are able to account for and model such dependencies, the consequences of which are significant. First, by allowing random effects to be part of the model, it is possible to not only identify but also to predict why differences between groups are occurring. Additionally, the groups, in this case, classrooms, are treated as a random sample from a larger population, allowing for inferences beyond the collected data (Hoffman, 2015).

The first step in developing the ultimate model is to create an unconditional or empty means, random intercept model. This model contains no predictors, thus empty, but does model the differences in the outcome variable, in this case NCE difference between grade eight and a grade six-seven average ($8 - 6/7avg$), thus random intercepts. This allows the residual variance in the outcomes, also seen in regression models, to be partitioned into the variation attributable to individual differences, residual, and to group differences, intercept. In education-based research
it is common to see relatively smaller proportions of variation attributable to the groups, with common values ranging from 0.05 to 0.20 (Hoffman, 2015). The empty means model for this study is presented below.

\[ \text{Level 1: } Y_{sc} = \beta_{0c} + \varepsilon_{sc} \]  
\[ \text{Level 2: } \beta_{0c} = \gamma_{00} + u_{0c} \]

Essentially a student \( s \) in class \( c \) has an NCE change \((8 - 6/7\text{avg})\) equal to the class average NCE change, \( \beta_{0c} \), plus some individual deviation of student \( s \), \( \varepsilon_{sc} \). That class average, \( \beta_{0c} \), is the result of the overall grade level average change, \( \gamma_{00} \), plus some group deviation of classroom \( c \). Putting the two levels into one equation yields the following MLM:

\[ Y_{sc} = \gamma_{00} + u_{0c} + \varepsilon_{sc} \]

\( u_{0c} \) and \( \varepsilon_{sc} \) represent the partitioned variance by group and individual, respectively. The results are shown in Table 10. This model provides a baseline model for fit and for partitioning the variance components. Two indicators of overall model fit will be included in each model summary table. The values for the -2 Restricted Log Likelihood (RLL) and Akaike’s Information Criterion (AIC) are statistical measures of fit. Overall, a smaller value indicates a better fit. For example, if the RLL presented in Table 10 of 1411.387 becomes smaller after the next predictor is included in the model, then the predictor improved the model fit—the greater the decrease, the greater the improvement in the model. MLMs function better when there are at least ten units at level two. This study only afforded eight level two classrooms or units, which may contribute to
the larger intercept variance $p$-value = 0.309. Nevertheless, this p-value need not be significant in the traditional sense to be important; the very nature of students being in classrooms warrants the utilization of an MLM.

Table 10

**Results for Empty Means Random Intercept Model**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>df</th>
<th>$t$</th>
<th>$p$</th>
<th>Information Criteria$^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercepts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-2 Restricted Log 1411.387</td>
</tr>
<tr>
<td>Estimates of Fixed Effects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Likelihood (RLL)</td>
</tr>
<tr>
<td>Intercept</td>
<td>3.470338</td>
<td>1.113906</td>
<td>7.003</td>
<td>3.115</td>
<td>.017</td>
<td>1415.387</td>
</tr>
<tr>
<td>Estimates of Covariance Parameters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Wald Z</td>
</tr>
<tr>
<td>Residual</td>
<td>104.830987</td>
<td>11.045467</td>
<td>9.491</td>
<td>.000</td>
<td>1415.387</td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>5.370094</td>
<td>5.274671</td>
<td>1.018</td>
<td>.309</td>
<td>1415.387</td>
<td></td>
</tr>
</tbody>
</table>

a. Dependent Variable: NCE 8 – 6/7avg.
b. The information criteria are displayed in smaller-is-better form.

From Table 10 the Intraclass Correlation (ICC) can be computed, a measure of how much of the variation in outcome is attributable to the nesting/grouping.

\[
\text{ICC} = \frac{\tau^2}{\tau^2 + \sigma^2} = \frac{5.37}{5.37 + 104.83} = 0.049
\]  

The model estimates nearly 5% of the variation in student relative achievement is attributable to the classrooms in which they were taught 8th grade ELA. Again, while this is on the low end of the spectrum, it is a reasonable value from which to explain additional group-based variance. Furthermore, the researcher also built other MLMs with the NCE from grade eight as the dependent variable. In that case, over 10% of the variation was attributable to students’ classrooms. However, of primary interest in this study was relative student achievement, and the way to assess that variable necessitated the change score (8 – 6/7avg) be the dependent variable.
Another way to illustrate the variation in 8\textsuperscript{th} grade classroom outcomes with respect to class period is to graph the overall change from students’ entering 6/7 average NCEs in a class and the class’s average NCE change score following the 8\textsuperscript{th} grade state assessment. Figure 4 represents this change graphically for each of the eight 8\textsuperscript{th} grade classes. The variation appears substantial.

\textit{Figure 4. Overall change in NCE by 8th grade class.}

Prior to continuing to build an MLM for this study, it was important to address the primary assumption required by MLMs: error values should be approximately normally distributed. Figure 5 shows the distribution of error terms from the empty means random intercept model and, as the distribution is approximately normal, verifies this assumption has been met.
Figure 5. The distribution of error terms/residuals by grouping sequence.

**Including Predictors in the MLM**

Knowing the requisite assumption of approximately normally distributed errors/residuals was met and the nesting data structure required an MLM, the next step was to add predictors to the model. To build the MLM, a process of forward selection was employed whereby predictors are systematically included and the resulting coefficients and standard errors are examined to determine whether or not to include them. The order of inclusion and checking was based in both the study’s hypotheses and the literature.

**Grouping sequence.** Based on the overall study hypothesis, the first predictor included was for grouping sequence.

\[
Y_{sc} = \beta_0c + \beta_1c X_{sc} + \epsilon_{sc} \quad (16)
\]

\[
\beta_0c = \gamma_00 + \mathbf{u}_{0c} \quad (17)
\]

\[
\beta_1c = \gamma_{10}
\]
The level of the predictor, \( X_{sc} \), is determined by the student, \( s \), and related to the class, \( c \). In this case, via dummy coding, the predictor takes on one of three possible values, non-advanced stay tracked (NAT), non-advanced to mixed (NAM), and advanced to mixed (AM). The results from including this predictor are shown in Table 11.

Table 11

*Results for Adding in Grouping Sequence Predictor*

<table>
<thead>
<tr>
<th>Source</th>
<th>Numerator df</th>
<th>Denominator df</th>
<th>( F )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>6.486</td>
<td>6.558</td>
<td>.040</td>
</tr>
<tr>
<td>Grouping Sequence</td>
<td>2</td>
<td>16.813</td>
<td>3.100</td>
<td>.071</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>df</th>
<th>( t )</th>
<th>( p )</th>
<th></th>
<th>Information Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.707744</td>
<td>1.620966</td>
<td>10.944</td>
<td>1.054</td>
<td>.315</td>
<td></td>
<td>Akaike's Information</td>
</tr>
<tr>
<td></td>
<td>.514307</td>
<td>3.047387</td>
<td>8.518</td>
<td>.169</td>
<td>.870</td>
<td></td>
<td>Criterion (AIC)</td>
</tr>
<tr>
<td>NAT</td>
<td>3.991038</td>
<td>1.643233</td>
<td>181.889</td>
<td>2.429</td>
<td>.016</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>Wald Z</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual</td>
<td>101.870250</td>
<td>10.770576</td>
<td>9.458</td>
<td>.000</td>
</tr>
<tr>
<td>Intercept</td>
<td>7.221954</td>
<td>6.724713</td>
<td>1.074</td>
<td>.283</td>
</tr>
</tbody>
</table>

Of note, when a predictor is included that has more than one (\( k \)) level, only \( k-1 \) of the levels will appear in the table with a coefficient; the last one appears as the intercept value. In this case, the level AM is estimated by the 1.707744 value in the table. Additionally, it is not immediately discernable from the estimates of fixed effects whether or not the predictor grouping sequence in the aggregate is significant as its components are separated. In this and other cases, an additional element is included in the table, the type III tests of fixed effects. These initial \( F \) statistics yield \( p \)-values for the overall impact of a predictor. From this overall test, the
intercept is significant \( p = 0.04 \), and the predictor grouping sequence is nearly as significant \( p = 0.07 \). Within the table, it is clear the difference between NAT and AM is insignificant due to NAT’s \( p \)-value of 0.87 and AM is the reference group. However, the difference between NAM and AM is significant \( p = 0.016 \). Considering this model to be parsimonious, it follows that AM, NAT, and NAM students are predicted to have average NCE gains of 1.71, 2.22 \((1.71 + 0.51)\), and 5.7 \((1.71 + 3.99)\), respectively. Lastly, this model is a better fit than the original due to both information criteria values, RLL and AIC, decreasing in value.

**Teacher sequence.** Based on the literature that teachers have a primary effect on student outcomes, the next predictor to be assessed was teacher sequence. Teacher sequence is a variable that attempts to combine a student’s two-year teacher effect into one four level variable. Teachers for 7\(^{th}\) and 8\(^{th}\) grades were also considered separately, but due to their large effects and interactions, the overall teacher sequence predictor principally accomplished the same goal. Over the course of two years of ELA, students had one of the following teacher combinations: AC, AD, BC, BD. For both clarity and brevity, the actual models are shown one more time.

\[
\begin{align*}
\text{Level 1 with two predictors } X: \quad Y_{sc} &= \beta_0c + \beta_{1c} X_{sc} + \beta_{2c} X_{sc} + \varepsilon_{sc} \tag{18} \\
\text{Level 2 slope equations:} \quad \beta_{0c} &= \gamma_{00} + u_{0c} \tag{19} \\
&\quad \beta_{1c} = \gamma_{10} \\
&\quad \beta_{2c} = \gamma_{20}
\end{align*}
\]

Thus far, the only random effect is for classroom average outcome, represented by \( \beta_{0c} \). It is possible to have other effects that vary across level two classrooms, meaning their effect on the outcome differs depending on the classroom. Neither grouping sequence nor teacher sequence is
appropriate to test for a random effect in that both contribute to the definition of the groups. Any Level 2 predictor, in a two level model, can only have a fixed effect (Hoffman, 2015). The results of also including teacher sequence as a predictor are provided in Table 12.

Table 12

Results for Adding in Teacher Sequence Predictor

<table>
<thead>
<tr>
<th>Type III Tests of Fixed Effectsa</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>Numerator df</td>
<td>Denominator df</td>
<td>F</td>
<td>p</td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>1</td>
<td>6.355</td>
<td>9.240</td>
<td>.021</td>
<td></td>
</tr>
<tr>
<td>Grouping Sequence</td>
<td>2</td>
<td>20.968</td>
<td>4.062</td>
<td>.032</td>
<td></td>
</tr>
<tr>
<td>Teacher Sequence</td>
<td>3</td>
<td>27.234</td>
<td>3.474</td>
<td>.030</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimates of Fixed Effectsa</th>
<th>Information Criteriab</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
<td>Estimate</td>
</tr>
<tr>
<td>Intercept</td>
<td>-1.657719</td>
</tr>
<tr>
<td>NAT</td>
<td>1.753561</td>
</tr>
<tr>
<td>NAM</td>
<td>4.630997</td>
</tr>
<tr>
<td>AC</td>
<td>6.858935</td>
</tr>
<tr>
<td>AD</td>
<td>2.347420</td>
</tr>
<tr>
<td>BC</td>
<td>1.611775</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimates of Covariance Parametersa</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual</td>
<td>Estimate</td>
<td>Std. Error</td>
<td>Wald Z</td>
<td>p</td>
</tr>
<tr>
<td></td>
<td>99.423552</td>
<td>10.559053</td>
<td>9.416</td>
<td>.000</td>
</tr>
<tr>
<td>Intercept</td>
<td>3.461753</td>
<td>4.721943</td>
<td>.733</td>
<td>.463</td>
</tr>
</tbody>
</table>

a. Dependent Variable: NCE 8 – 6/7avg.
b. The information criteria are displayed in smaller-is-better form.

From the type III tests, the addition of teacher sequence as a predictor is significant and decreased the p-values of the overall intercept and grouping sequence predictor. Often with MLMs, as additional predictors are included, the significance of other predictors change, sometimes substantially. Again, for the teacher sequence variable only three of the four values are directly given coefficients. In this case, sequence BD is captured within the intercept value, which now includes information for students who were advanced and moved to mixed ability.
classes (AM) who also had teacher B in 7th grade and teacher D in 8th grade (BD). The model estimates this combination, on average, leads to a decrease in NCE units of 1.66, though not in a significant way ($p = 0.415$). Essentially, the tendency is to have a small decline with this combination; however, the variation in outcomes is large.

As before, BD serves as the reference group against which the other three teacher combinations are compared and situated. For instance, AC has a coefficient of 6.86, indicating students experiencing this combination of teachers, on average, increased by 6.86 more NCE points than students having teachers BD. This increase is statistically significant ($p = 0.014$), indicating a distinct benefit of having teachers AC over teachers BD. The other direct comparisons against BD are not significant and it is also not directly possible to compare other sequences to one another, such as AC to BC. This information could be obtained by altering the reference group, and, while potentially interesting, is not of interest to this study, not only due to the primary interest in grouping sequence relationships, but also due to the overall interest in determining significance in general so as to allow inferences to extend to the general population of teacher sequences. As such, even if teacher effects were of consequence to this study, specific pairwise differences would not necessarily be important unless other predictors could contribute to the explanation of those differences.

In combining the grouping sequence effect with the teacher sequence effect, the model predicts a profound impact for non-advanced students who also had teacher sequence AC. Both coefficients are statistically significant ($p = 0.005, 0.014$) and large (4.63, 6.86). Their joint effect after including the intercept correction is 9.83NCEs. An increase of nearly ten NCE units over the course of one year of instruction would be remarkable. The inclusion of teacher sequence improved the model as noted by the 20 point decreases in both RLL and AIC and by
the changes in the covariance parameter estimates. Including teacher sequence decreased the residual variance component by about two percent. However, the intercept variance decreased by 52%, indicating that of the variation in student outcomes attributable to students being in different classrooms, over half of that variation is due to the teachers themselves.

**Race.** Racial distribution in tracked classes was a motivating factor to this study and is based in the literature as a related element of tracking students by ability; therefore, race was the next predictor examined. In this study, race took on only three values, White, African-American, and Latino. While some students in the cohort were of a different race, no single other subgroup substantiated a large enough group from which to determine relationships or effects. Putting them all into one group labeled “Other” would also have been too small of a group and would have conflated results within the group. As such, only White, African-American, and Latino students were included for this portion of the analysis. The results are presented in Table 13.
Table 13

Results for Race as a Predictor

<table>
<thead>
<tr>
<th>Source</th>
<th>Numerator df</th>
<th>Denominator df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>7.410</td>
<td>11.388</td>
<td>.011</td>
</tr>
<tr>
<td>Grouping Sequence</td>
<td>2</td>
<td>26.476</td>
<td>4.001</td>
<td>.030</td>
</tr>
<tr>
<td>Teacher Sequence</td>
<td>3</td>
<td>36.292</td>
<td>4.380</td>
<td>.010</td>
</tr>
<tr>
<td>Race</td>
<td>2</td>
<td>164.097</td>
<td>1.024</td>
<td>.361</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>df</th>
<th>t</th>
<th>p</th>
<th>Information Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.707592</td>
<td>2.712967</td>
<td>57.368</td>
<td>-.629</td>
<td>.532</td>
<td>-2 Restricted Log</td>
</tr>
<tr>
<td>NAT</td>
<td>2.470378</td>
<td>2.571725</td>
<td>13.161</td>
<td>.961</td>
<td>.354</td>
<td>Likelihood (RLL)</td>
</tr>
<tr>
<td>NAM</td>
<td>5.669481</td>
<td>2.031504</td>
<td>164.363</td>
<td>2.791</td>
<td>.006</td>
<td>Akaike's Information</td>
</tr>
<tr>
<td>AC</td>
<td>7.849939</td>
<td>2.236916</td>
<td>13.718</td>
<td>3.509</td>
<td>.004</td>
<td>Criterion (AIC)</td>
</tr>
<tr>
<td>AD</td>
<td>2.645471</td>
<td>2.133986</td>
<td>164.950</td>
<td>1.240</td>
<td>.217</td>
<td></td>
</tr>
<tr>
<td>BC</td>
<td>2.328865</td>
<td>2.437132</td>
<td>19.099</td>
<td>.956</td>
<td>.351</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>-.387088</td>
<td>2.402682</td>
<td>164.789</td>
<td>-.161</td>
<td>.872</td>
<td></td>
</tr>
<tr>
<td>AfricAmer</td>
<td>-2.435970</td>
<td>1.929862</td>
<td>162.898</td>
<td>-1.262</td>
<td>.209</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>Wald Z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual</td>
<td>98.653084</td>
<td>11.004397</td>
<td>8.965</td>
<td>.000</td>
</tr>
<tr>
<td>Intercept</td>
<td>1.749612</td>
<td>3.750754</td>
<td>.466</td>
<td>.641</td>
</tr>
</tbody>
</table>

Including the predictor race in the MLM did not significantly alter the other predictors’ effects, indicating a lack of interaction between race and both teacher and grouping sequences. The collective significance of race as a level one student outcome predictor was $p = 0.361$, which is not statistically significant, though one could argue it is trending that way, particularly with a limited sample. Ultimately, in determining whether or not to maintain race as a model predictor, one should also consider possible significant reference group contrasts shown in the fixed effects estimates. In this case, Latino served as the reference group, with results more similar than different with respect to White ($p = 0.872$) and more different than similar to African American,
though not significantly so \((p = 0.209)\). Another indicator of a predictor’s worth to a model can be derived from both the effect on the variance components and on the RLL and AIC. In both cases, race contributes to the model in important ways. Both the RLL and AIC decreased by an appreciable 123 points, indicating substantial improvement in model fit.

Interestingly, while race is an individual attribute, almost no additional residual variance was explained by including race as a level one predictor. However, the intercept variance, which is a group level component, again decreased by 50%. Such a result signals race should also be aggregated to a level two predictor. Yet when the level two race variable was also included, no change occurred. Additionally, race was entered as a random effect to determine if how race affected average outcomes varied across classrooms. While the model converged, the Hessian matrix was not positive definite. Hessian matrices are utilized with MLMs to estimate the variances and covariances associated with random effects. Convergence without a positive definite Hessian matrix typically means the estimated parameter values are reliable, but the associated standard errors are not due to an insufficient amount of observed variation. As such, statistical significance cannot be confirmed. From the results it appeared race was not a significant random effect, though with the Hessian warning, it cannot be guaranteed. Therefore, for the purposes of this analysis, it was assumed the entire effect of race was captured within the level one predictor, which will stay in the model.

**Sex.** To assess whether differential relationships existed between male and female students, sex was next entered into the model. The results are presented in Table 14.
Table 14

*Results for Including Sex as a Predictor*

<table>
<thead>
<tr>
<th>Source</th>
<th>Numerator df</th>
<th>Denominator df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>13.004</td>
<td>4.022</td>
<td>.066</td>
</tr>
<tr>
<td>Grouping Sequence</td>
<td>2</td>
<td>25.803</td>
<td>3.441</td>
<td>.047</td>
</tr>
<tr>
<td>Teacher Sequence</td>
<td>3</td>
<td>38.366</td>
<td>5.332</td>
<td>.004</td>
</tr>
<tr>
<td>Race</td>
<td>2</td>
<td>163.359</td>
<td>.873</td>
<td>.420</td>
</tr>
<tr>
<td>Sex</td>
<td>1</td>
<td>163.923</td>
<td>5.361</td>
<td>.022</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>df</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.600772</td>
<td>2.664284</td>
<td>67.094</td>
<td>-.976</td>
<td>.332</td>
</tr>
<tr>
<td>NAT</td>
<td>2.377468</td>
<td>2.439178</td>
<td>14.193</td>
<td>.975</td>
<td>.346</td>
</tr>
<tr>
<td>NAM</td>
<td>5.217101</td>
<td>2.011850</td>
<td>161.512</td>
<td>2.593</td>
<td>.010</td>
</tr>
<tr>
<td>AC</td>
<td>8.184241</td>
<td>2.127768</td>
<td>15.069</td>
<td>3.846</td>
<td>.002</td>
</tr>
<tr>
<td>AD</td>
<td>2.809512</td>
<td>2.106087</td>
<td>163.628</td>
<td>1.334</td>
<td>.184</td>
</tr>
<tr>
<td>BC</td>
<td>1.757871</td>
<td>2.343062</td>
<td>21.649</td>
<td>.750</td>
<td>.461</td>
</tr>
<tr>
<td>White</td>
<td>-.740558</td>
<td>2.374312</td>
<td>162.853</td>
<td>-.312</td>
<td>.756</td>
</tr>
<tr>
<td>AfricAmer</td>
<td>-2.357771</td>
<td>1.908141</td>
<td>162.054</td>
<td>-1.236</td>
<td>.218</td>
</tr>
<tr>
<td>Sex</td>
<td>3.892125</td>
<td>1.681006</td>
<td>163.923</td>
<td>2.315</td>
<td>.022</td>
</tr>
</tbody>
</table>

As seen in the table, sex is a significant contributor to the model even after conditioning for all prior predictors (p = 0.022). For this variable, zero indicated a female student and one a male student. Hence, females served as the reference group and were absorbed in the intercept, while the effect of being male is included directly in the table as a coefficient of 3.89. This means the model estimates between 7th and 8th grade, male students would, on average, gain 3.89 more NCE points on their ELA state assessment. Tracking with all the reference groups so far, the intercept indicates the expected average gain of a female Latina advanced student who had
teacher B in 7th grade and teacher D in 8th grade. For this specific combination of predictors, an average loss of 2.6 NCE points is predicted.

The low $p$-value for sex means it should remain in the model. Examining further, both indicators of model fit, RLL and AIC, improved by about eight points. Again the intercept variance decreased by almost 46%; however, in that the amount of variation at level two, the intercept, has become so small, this may or may not indicate the need to include sex as a level two predictor. The residual variance decreased a small amount and remains significant. To be thorough, sex will be entered as a level two predictor and grand-mean centered. Essentially that means the value of the predictor will be the difference between the proportion of the class that is male and the overall population proportion male, 0.31. Grand mean centering is one way to force a predictor to have a meaningful value of zero, which aids in interpretation (Hoffman, 2015; Raudenbush & Bryk, 2002). In that sex is a binary predictor at level one, no centering was necessary, as zero already meant female. Technically, at level two, zero would still have meaning, though would mean a class of 0% males, which did not occur. Therefore, subtracting the grand mean, in this case 0.31, centers the values, so a zero value would mean the class was 31% male. All predictors prior to this point were categorical, and therefore the zero value having meaning was addressed via dummy coding. For instance, in this study grouping sequence was one predictor with three levels. Each level became its own variable, which is why NAT and NAM show up in the table of fixed effects with their own coefficients. The third level, AM, is absorbed in the intercept. So, to predict a value for an AM student, zeros would be entered in the NAT and NAM predictors, negating their overall effect entirely. The results for including the grand-mean centered level two sex predictor are presented in Table 15.
As shown in the information criteria portion of the table, including the level two predictor for sex decreased both the RLL and AIC by around seven points, from approximately 1247 and 1251 to 1240 and 1244, respectively. Such a decrease indicates including the class level proportion of students who were male as a predictor improved the overall fit of the model. The predictor itself only had a $p$-value of 0.487 even though conditioning for proportion male improved the general fit. The fixed effect coefficients and $p$-values did not alter much, while the
variance components actually slightly increased. In that sex was such a significant level one predictor representing a student’s individual sex, and the overall model improved by including sex at level two, representing the proportion of a class that was male, both will remain in the model.

**Prior Ability Indicators.** While the dependent variable of change in NCE score from one year to the next would not seem to depend on initial ability, due to the hypothesis that non-advanced students would perform better in mixed ability settings, checking for an effect of ability was prudent. As such, student’s group-mean centered average NCE score from grades six and seven was entered as the next predictor. Group mean centering is another type of centering available for use at level one. Again the goal is to have zero mean something, and in this case, zero would mean the student scored at the average value of their 8th grade classroom. It is also acceptable to grand mean center level one variables. However, the benefit of group-mean centering a level one predictor is that by centering the values within their respective groups, all level two information is effectively stripped from the data, creating a predictor and resultant coefficients that only pertain to level one. In this instance, the level two predictor associated with prior ability was simultaneously entered into the model. Decisions as to adding predictors at one level at a time or together are empirical in nature (Hoffman, 2015). In this case, as both address the same construct of initial ability, the two predictors were entered jointly. The level two predictor was grand mean centered, meaning the overall mean from all students’ grade six and seven NCE average was subtracted from the class average of the students’ collective grade six and seven NCE average. It follows an 8th grade class of lower entering ability would have a negative value while a higher ability class would have a positive value for the predictor. The results of conditioning for initial ability at both levels one and two are provided in Table 16.
Table 16

*Results for Including Initial Ability as Predictors at Both Levels One and Two*

<table>
<thead>
<tr>
<th>Source</th>
<th>Numerator df</th>
<th>Denominator df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>4.981</td>
<td>.691</td>
<td>.444</td>
</tr>
<tr>
<td>Grouping Sequence</td>
<td>2</td>
<td>6.534</td>
<td>.783</td>
<td>.496</td>
</tr>
<tr>
<td>Teacher Sequence</td>
<td>3</td>
<td>13.567</td>
<td>6.486</td>
<td>.006</td>
</tr>
<tr>
<td>Race</td>
<td>2</td>
<td>160.106</td>
<td>2.063</td>
<td>.130</td>
</tr>
<tr>
<td>Sex-Individual</td>
<td>1</td>
<td>159.960</td>
<td>5.756</td>
<td>.018</td>
</tr>
<tr>
<td>SexClassLevel</td>
<td>1</td>
<td>3.549</td>
<td>1.451</td>
<td>.303</td>
</tr>
<tr>
<td>Initial Ability-Ind</td>
<td>1</td>
<td>160.490</td>
<td>18.993</td>
<td>.000</td>
</tr>
<tr>
<td>Initial Ability-Class</td>
<td>1</td>
<td>3.388</td>
<td>1.648</td>
<td>.280</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>df</th>
<th>t</th>
<th>p</th>
<th>Information Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>.101887</td>
<td>2.712473</td>
<td>55.298</td>
<td>.038</td>
<td>.970</td>
<td>Restricted Log 1224.679</td>
</tr>
<tr>
<td>NAT</td>
<td>-9.237041</td>
<td>7.386758</td>
<td>4.125</td>
<td>-1.250</td>
<td>.277</td>
<td>Likelihood (RLL)</td>
</tr>
<tr>
<td>NAM</td>
<td>-.862230</td>
<td>2.384325</td>
<td>160.752</td>
<td>-.362</td>
<td>.718</td>
<td>Akaike's Information 1228.679</td>
</tr>
<tr>
<td>AC</td>
<td>9.809233</td>
<td>2.338790</td>
<td>5.893</td>
<td>4.194</td>
<td>.006</td>
<td>Criterion (AIC)</td>
</tr>
<tr>
<td>AD</td>
<td>3.748911</td>
<td>2.035511</td>
<td>159.899</td>
<td>1.842</td>
<td>.067</td>
<td></td>
</tr>
<tr>
<td>BC</td>
<td>3.642548</td>
<td>2.518745</td>
<td>8.085</td>
<td>1.446</td>
<td>.186</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>2.091395</td>
<td>2.386604</td>
<td>160.533</td>
<td>.876</td>
<td>.382</td>
<td></td>
</tr>
<tr>
<td>AfricAmer</td>
<td>-1.856242</td>
<td>1.828952</td>
<td>158.638</td>
<td>-1.015</td>
<td>.312</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>3.892125</td>
<td>1.681006</td>
<td>163.923</td>
<td>2.315</td>
<td>.022</td>
<td></td>
</tr>
<tr>
<td>SexClassLevel</td>
<td>11.308626</td>
<td>9.389647</td>
<td>3.549</td>
<td>1.204</td>
<td>.303</td>
<td></td>
</tr>
<tr>
<td>Intl Ability-Ind</td>
<td>-.254132</td>
<td>.058313</td>
<td>160.490</td>
<td>-4.358</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Intl Ability-Cls</td>
<td>-.704164</td>
<td>.546382</td>
<td>3.388</td>
<td>-1.284</td>
<td>.280</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>Wald Z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual</td>
<td>87.498824</td>
<td>9.830738</td>
<td>8.901</td>
<td>.000</td>
</tr>
<tr>
<td>Intercept</td>
<td>.035620</td>
<td>3.175150</td>
<td>.011</td>
<td>.991</td>
</tr>
</tbody>
</table>

---

a. Dependent Variable: NCE 8 – 6/7avg.
b. The information criteria are displayed in smaller-is-better form.

Including both predictors for initial ability as measured by the average grade six and grade seven NCE score decreased the model fit indicators by a substantial 23 points. From the type III tests, individual initial ability was highly significant ($p < 0.000$), while class ability level
was less so, having a $p$-value of 0.280, and the intercept $p$-value changed considerably. Similarly, a number of fixed effect estimates and their corresponding $p$-values also changed substantially after conditioning for both individual and class initial ability. Perhaps the more specific initial ability predictors are accounting for more of the variation than the more dilute ability predictors captured within the grouping sequence variable. Lastly, the residual variance component lowered by about ten percent, the largest decrease so far, and the intercept variance all but disappeared. As such, both initial ability predictors will remain in the model.

**Support Class.** After seeing the significant impact the initial ability predictors had on the model, a final predictor to condition for the most at risk students was included. A number of students were in one of three types of support classes: a reading remediation class designed specifically to support struggling readers, a learning strategies class to support students with disabilities, or a class supporting English language learners. In each case, the support class served as one of the three electives students had during their 8th grade year. In that initial ability related significantly to NCE gains in 8th grade, perhaps support classes also contributed to gains even though students had the same supports in the prior year. Similar to the predictor sex, support class had two levels, with zero indicating the student did not have a support class and one indicating they did. Therefore, it was not necessary to center the level one predictor. However, the level two predictor for support class was grand mean centered. 19.4% of the students in the study had a support class during 8th grade, so the proportion of those students in each 8th grade class was centered around 0.194 and included as a level two predictor. The results of including both predictors to condition for whether or not students were enrolled in a support class are included in Table 17.
Table 17

Results for Including Having a Support Class as a Predictor at Levels One and Two

<table>
<thead>
<tr>
<th>Source</th>
<th>Numerator df</th>
<th>Denominator df</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>3.347</td>
<td>1.638</td>
<td>.282</td>
</tr>
<tr>
<td>Grouping Sequence</td>
<td>2</td>
<td>3.722</td>
<td>.693</td>
<td>.555</td>
</tr>
<tr>
<td>Teacher Sequence</td>
<td>3</td>
<td>6.036</td>
<td>5.450</td>
<td>.037</td>
</tr>
<tr>
<td>Race</td>
<td>2</td>
<td>157.565</td>
<td>2.928</td>
<td>.056</td>
</tr>
<tr>
<td>Sex-Individual</td>
<td>1</td>
<td>157.780</td>
<td>7.501</td>
<td>.007</td>
</tr>
<tr>
<td>SexClassLevel</td>
<td>1</td>
<td>2.258</td>
<td>.945</td>
<td>.423</td>
</tr>
<tr>
<td>Initial Ability-Ind</td>
<td>1</td>
<td>158.595</td>
<td>25.619</td>
<td>.000</td>
</tr>
<tr>
<td>Initial Ability-Class</td>
<td>1</td>
<td>2.617</td>
<td>.966</td>
<td>.408</td>
</tr>
<tr>
<td>Support Class-Ind</td>
<td>1</td>
<td>157.379</td>
<td>6.410</td>
<td>.012</td>
</tr>
<tr>
<td>Support Class-Class</td>
<td>1</td>
<td>2.396</td>
<td>.024</td>
<td>.890</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>df</th>
<th>t</th>
<th>p</th>
<th>Information Criteria</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.485345</td>
<td>2.832785</td>
<td>27.174</td>
<td>.524</td>
<td>.604</td>
<td>-2 Restricted Log</td>
<td>1207.238</td>
</tr>
<tr>
<td>NAT</td>
<td>-10.22264</td>
<td>8.755402</td>
<td>2.786</td>
<td>-1.168</td>
<td>.333</td>
<td>Likelihood (RLL)</td>
<td></td>
</tr>
<tr>
<td>NAM</td>
<td>-1.173184</td>
<td>2.371204</td>
<td>158.628</td>
<td>-.495</td>
<td>.621</td>
<td>Akaike's Information</td>
<td>1211.238</td>
</tr>
<tr>
<td>AC</td>
<td>10.185433</td>
<td>3.066072</td>
<td>3.682</td>
<td>3.322</td>
<td>.033</td>
<td>Criterion (AIC)</td>
<td></td>
</tr>
<tr>
<td>AD</td>
<td>3.582712</td>
<td>2.012323</td>
<td>158.359</td>
<td>1.780</td>
<td>.077</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BC</td>
<td>3.614051</td>
<td>3.170461</td>
<td>4.333</td>
<td>1.140</td>
<td>.313</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>2.338821</td>
<td>2.363968</td>
<td>157.539</td>
<td>.989</td>
<td>.324</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AfricAmer</td>
<td>-2.334050</td>
<td>1.810891</td>
<td>157.395</td>
<td>-1.289</td>
<td>.199</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>4.425345</td>
<td>1.615766</td>
<td>157.780</td>
<td>2.739</td>
<td>.007</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SexClassLevel</td>
<td>10.510171</td>
<td>10.813859</td>
<td>2.258</td>
<td>.972</td>
<td>.423</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intl Ability-Ind</td>
<td>-.332528</td>
<td>.065697</td>
<td>158.595</td>
<td>-5.062</td>
<td>.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intl Ability-Class</td>
<td>-.802960</td>
<td>.817099</td>
<td>2.617</td>
<td>-.983</td>
<td>.408</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sprt Class-Ind</td>
<td>-5.588469</td>
<td>2.207238</td>
<td>157.379</td>
<td>-2.532</td>
<td>.012</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sprt Class-Class</td>
<td>3.048614</td>
<td>19.776464</td>
<td>2.396</td>
<td>.154</td>
<td>.890</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>Wald Z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual</td>
<td>84.752232</td>
<td>9.558523</td>
<td>8.867</td>
<td>.000</td>
</tr>
<tr>
<td>Intercept</td>
<td>1.471077</td>
<td>5.155819</td>
<td>.285</td>
<td>.775</td>
</tr>
</tbody>
</table>

a. Dependent Variable: NCE 8 – 6/7avg.
b. The information criteria are displayed in smaller-is-better form.
Including both support class predictors improved the overall fit of the model, decreasing both the RLL and AIC by 17 points. The individual predictor for support class was significant ($p = 0.012$), though the class level predictor was not ($p = 0.890$). The residual variance decreased by approximately 3%, while the intercept variance increased substantially. Putting together the RLL and AIC decrease, the significance levels of both support class predictors, the shifts in variance, and the goal of a parsimonious model related to the hypotheses of the study, the final model will not include either support class variables. The model collapsed via the Hessian matrix non-positive definite factor when the individual support class predictor was included without the class level predictor. Otherwise, due to its individual significance ($p = 0.012$), the support class Level 1 predictor would have been included. The final model will be further interpreted in the following section.

**Interpreting the Final Multilevel Model**

The final model presented in Table 16 includes 14 parameter estimates to effectively model the change in NCE score from a baseline grade six/seven average to the 8th grade outcome from the various predictors. The individual student level predictors at Level 1 include the time-invariant characteristics of race and sex, of which sex remained statistically significant in the final model. One other student level predictor reflected a characteristic more prone to vary over time, an indicator of student ability, as measured by state test performance. This predictor was the most statistically significant of any in the model ($p < 0.000$).

Of the four classroom level predictors, only teacher sequence remained statistically significant ($p = 0.006$) in the final model. Grouping sequence (0.496), the proportion of males in a class (0.303), and the collective initial ability of a class (0.280) all had insignificant $p$-values in the final model. All remained part of the final model due to their relationship with Level 1
predictors or to their prior predictive significance. Of note, a primary consideration of this study, grouping sequence, had a p-value of 0.496 in the final model. However, grouping sequence remained a statistically significant part of the model up until predictors that would condition for student and class initial ability were included. To some extent, grouping sequence and initial ability are conflated in that two of the grouping trajectories proxy as gross indicators for lower ability students and the third trajectory loosely captures predictive power for higher ability students. To gain a clearer sense of what the model is indicating with respect to how ability is predicted to relate to 8th grade relative NCE achievement, consider a student who started their 8th grade year with an average six/seven NCE ten points below their class’ average. That student’s value for the initial ability predictor would be -10, which would be multiplied by the model coefficient -0.254 to yield that individual student’s initial ability effect. The model predicts such students, on average and absent any other conditions, would have a gain of 2.45 NCE points during their 8th grade year.

It is important to remember several reference groups are collectively captured by the intercept value of 0.102, namely females, Latinos, students jointly with teachers B and D, and advanced students who moved to mixed level ELA classes in 8th grade. Following the sequence of predictors as presented in Table 16, the final MLM is provided both symbolically and with the estimated coefficients. The model uses all the individual student predictor values to predict a change in NCE score for each student. Recall the $u_{0c}$ and $\varepsilon_{sc}$ represent the error or correction terms for both class and individual that move a student’s predicted value to his or her actual value. Additionally, all $\gamma_{s0}$ values are constant due to the model supporting just one random effect.
Symbolic representation of the final MLM:

\[ Y_{sc} = \gamma_0 + \gamma_{10} X_{sc1} + \gamma_{20} X_{sc2} + \gamma_{30} X_{sc3} + \gamma_{40} X_{sc4} + \gamma_{50} X_{sc5} + \gamma_{60} X_{sc6} + \gamma_{70} X_{sc7} + \gamma_{80} X_{sc8} + \gamma_{90} X_{sc9} + \gamma_{100} X_{sc10} + \gamma_{110} X_{sc11} + u_0 + \epsilon_{sc} \]  

(20)

Actual MLM with inserted coefficients:

\[ Y_{sc} = 0.102 + -9.237 X_{sc1} + -0.862 X_{sc2} + 9.809 X_{sc3} + 3.749 X_{sc4} + 3.643 X_{sc5} + 2.091 X_{sc6} + -1.856 X_{sc7} + 3.892 X_{sc8} + 11.309 X_{sc9} + -0.254 X_{sc10} + -0.701 X_{sc11} + u_0 + \epsilon_{sc} \]  

(21)

To further illuminate how the model effectively produces predictions, consider the previously discussed ten point below average student along with the following additional information: the student is male, in a class that is 41% male, his class’ initial class ability was the same as the overall grade level average, he is African American, had teachers BD, and was a non-advanced student in a mixed level class in 8th grade. Below is how all this individual student’s descriptive information comes together in the model, following the same order presented:

\[
\text{Student gain prediction} = -0.254(-10) + 3.893(1) + 11.309(0.10) + -0.701(0) + -1.856(1) + 0.102 + -0.862(1) = 4.9479
\]  

(22)

Altogether then, the previously described student would be predicted to have an NCE gain of nearly 5 points. Note for predictors in which either the student had the characteristic or
not, a simple one or zero determines the value, while the three centered predictors’ input values had to first be calculated as the individual’s or class’ value minus the centered value. Again, the results from the MLM are only predicted gains and some of the predictors include substantial variation. It can be more instructive to pay closer attention to how the statistically significant predictors of teacher sequence, race, sex, and initial ability interact with one another. Additional exploration of the coefficients and their implications for this study’s hypotheses will be presented in Chapter 5.

**Rechecking the Assumption of Normally Distributed Residuals**

Following the establishment of the final model, it was important to again examine the distribution of the residuals. The normality of the residuals, a requirement for MLMs, was assessed both graphically and by way of the Shapiro-Wilk test. The Shapiro-Wilk test examines the distribution of residuals against a null hypothesis assuming the population from which the data came is normal. The test statistic, $W(173)$, was insignificant ($p = 0.836$), indicating there was no evidence to suggest the residuals originated from a non-normal population. Figure 6 provides the visual inspection for normality, which is in accord with the Shapiro-Wilk test.
Figure 6. Distribution of residuals from the final MLM.

Lastly, it is standard procedure to check a distribution’s normality with a normal Q-Q plot.

Essentially, if the data points are concentrated around the line, the data are assumed to be from a normal population. Figure 7 presents the normal Q-Q plot of the final MLM’s residual, again confirming the normality of the data.
Figure 7. Normal Q-Q plot of the final MLM’s residuals.

Summary

In this chapter, relevant descriptive statistics were presented to both add contextual value to the study and to make preliminary assessments regarding the study’s hypotheses. The motivations for and results of a number of preliminary tests, such as Chi-square and matched pairs, were presented and discussed. The need for a more complex analytical technique, multilevel modeling, was offered and followed. For the development of the MLM, the inclusion or exclusion of each predictor at level one and level two was sequentially presented until the model was finalized. Brief explanations of the actual meaning of the model and how it functioned were presented, ultimately culminating with one complete example and interpretation. In the next chapter, the results will be further discussed, including implications for policy, practice, and leadership preparation. Additionally, limitations of the study will be noted along with suggestions for future research.
CHAPTER FIVE: DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS

Introduction

The purpose of this study was to investigate the relationship between student to classroom assignment practices for English Language Arts (ELA) instruction and resultant student achievement on North Carolina End of Grade (EOG) assessments. The investigation took place in an urban magnet middle school that afforded a unique opportunity to study data from students who were tracked by ability in 7th grade then assigned to mixed level ELA classes in 8th grade. Additionally, 18% of the students from the same cohort with the same teachers remained in lower ability tracked classrooms in 8th grade. Therefore, in addition to being able to examine students’ individual change in EOG performance from 7th to 8th grade that coincided with a change in classroom assignment, contrasts across the three grouping trajectories were examined within one predictive multilevel model (MLM). The three trajectories were non-advanced students moving from low tracked 7th grade ELA classes to mixed 8th grade ELA classes, advanced students who were in high tracked classes in 7th and mixed classes in 8th grade, and non-advanced students who were in low tracked classes in 7th and remained in low tracked classes in 8th grade. As such, the following hypotheses guided the study:

Major Research Hypothesis

How students are assigned to classes in the secondary school setting has a relationship with relative student achievement on standardized state tests.
Additional Research Hypotheses

**Sub-Hypothesis 1:** Non-advanced students have better relative achievement when they are assigned to mixed level classes than when they are in tracked classes.

**Sub-Hypothesis 2:** Advanced students’ relative achievement is independent of the type of class to which they are assigned.

**Sub-Hypothesis 3:** Mixed level classrooms better reflect the school’s racial distribution than tracked classrooms.

The major hypothesis directly relates to the first two sub-hypotheses. For example, if non-advanced students were found to have better relative achievement when assigned to mixed level classes than when they are in tracked classes, there would be evidence to support the major hypothesis. On the other hand, if advanced students’ relative achievement were found to be independent of their class assignment, then no evidence would be added to support the major hypothesis, except in how it interacts with the findings from the first sub-hypothesis.

Additionally, if advanced students were found to have relative achievement related to the type of class assignment practice, then the major research hypothesis would be supported. In that the support or lack of support of the first two sub-hypotheses directly affects the support for the major hypothesis, the evidence from the study relating to the first two sub-hypotheses will be the primary discussion points in this chapter, along with evidence to support or discredit sub-hypothesis three.

**Interpretation of the Results**

In this section, results presented in Chapter 4 will be revisited and interpreted with respect to the three sub-hypotheses of interest in the study. For each sub-hypothesis, descriptive statistics, inferential tests, and multilevel model components will be offered in specific relation
to the particular hypothesis. Finally, results from the first two sub-hypotheses will be synthesized and proposed as evidence in support of or against the major research hypothesis of the study.

With regard to MLM-based results, it is important to note resultant $p$-values are conservative. This is due to the fact that MLMs address both outcome dependence resulting from students being nested within classrooms and aggregation bias via larger standard error estimates (Raudenbush & Bryk, 2002; Singer & Willett, 2003).

**Results Related to Sub-Hypothesis 1**

The study hypothesized non-advanced students would perform better on culminating standardized state tests when they were assigned to mixed ability classes as opposed to being assigned to tracked classes with only non-advanced students. Much of the literature around tracking and student achievement has focused on the deleterious effects of tracking on lower level students (see, for example, Ansalone, 2003; Archbald & Keleher, 2008; Carbonaro, 2005; Lleras & Rangel, 2009; Oakes, 2005). The literature is less conclusive regarding the benefits of tracking for higher ability students, though a number of studies have demonstrated benefits of grouping gifted students together for instruction (Goldring, 1990; Rogers, 1993; Rubin, 2003). One of the primary talking points in education for the past several decades has been the achievement gap. Interestingly, in spite of the consistency in the literature regarding the differential effects of tracking and its role in widening the achievement gap, the practice nonetheless persists in approximately 80 to 85% of American secondary schools (Ansalone, 2010; Archbald et al., 2009; Kozol, 2006; Loveless, 2013; Oakes, Gamoran, & Page, 1992).

The aim of this study was not to compare gains of low tracked students to those of high tracked students, which has been the analysis presented in most related studies, but rather to compare students who were in low tracked classrooms in 7th grade to the same students in mixed
ability 8th grade classrooms. Studies have attempted to make such comparisons across schools by matching students at one school with similar students at another school. However, in these studies a host of other variables are presented that need to be controlled, such as school and geographic residential effects. In this study, students could be compared with themselves and within the same school across two different years that coincided with either being tracked or mixed by ability for ELA instruction.

The non-advanced students who moved into mixed ability ELA classes in 8th grade had a statistically significant average NCE gain of 5.51 points in the aggregate and 4.70 NCE points for individual student matched differences. As this is the actual study result, one can predict how the population of non-advanced students moving into mixed classes would perform. With 95% confidence, the true one year average expected NCE gain is between 3.24 and 7.77 points. Perhaps all students at this school improved and the change had nothing to do with their change in class assignment. Indeed, their non-advanced peers also experienced an average NCE gain in 8th grade of 1.96 points in the aggregate and 1.06 NCE points for individual matched differences. However, not only are their gains smaller than their mixed ability peers, they also are insignificant, meaning this group of non-advanced students who remained in low track classes could just as easily have had no gain or a loss. The 95% confidence interval for this group of students includes anywhere from an average loss of 1.80 to a gain of 5.72 NCE points.

The descriptive statistics, including the ability to utilize matched pairs tests to control for individual student characteristics, clearly support the sub-hypothesis that non-advanced students perform better when assigned to mixed ability classes than when they are in ability tracked classes. However, as presented in both Chapters 3 and 4, when students are nested within classrooms, with different groups of peers, and taught by different teachers at different times of
the day, it is prudent to check for significant fixed or random effects of the predictors of interest within a multilevel model (MLM). In the parsimonious model that controlled for class level variation, a joint effect of peers and 8th grade teacher, via a random intercept, the overall effect of grouping sequence as a predictor of change score was $p = 0.071$. Within the model, the fixed effect for non-advanced students moving to mixed classes had the highest grouping effect, 3.99, and statistically different from advanced students in mixed classes ($p = 0.016$). Essentially, in this simplest form, sub-hypothesis one is supported, indicating non-advanced students were predicted to gain an average of 5.70 (3.99 + 1.71) NCE points, while their counterpart non-advanced peers who remained tracked were predicted to gain 2.22 (0.51 + 1.71) points, and the advanced to mixed students were predicted to maintain their prior NCE level.

Importantly, beginning with the first additional predictor, as variables were sequentially added to the model, the overall effect of grouping sequence became statistically significant and ranged from $p = 0.030$ to $p = 0.047$ until initial ability student- and class-level controls were included in the MLM. At that point the $p$-value jumped to 0.496. However, this $p$-value is not a direct indicator of grouping sequence’s worth as a predictor of NCE change. Rather, with the inclusion of predictors, it becomes a conditional indicator of an effect, meaning it expresses the significance of grouping after other variables have been controlled or conditioned. Furthermore, the substantial change in $p$-value when initial ability controls were included was not surprising in that initial ability played a large role in determining a student’s grouping sequence. The significant ($p < 0.000$) coefficient for group-centered initial student ability was negative, meaning for students of below average ability, their predicted gain would be positive. Most non-advanced students in mixed classrooms began with a below average initial ability when compared to the whole class average, so their negative initial ability deviation multiplied by
-0.254 becomes a positive gain prediction. When combined with their insignificant -0.862 coefficient for being non-advanced in a mixed class, the result will almost always continue to yield a positive gain prediction, not to mention the consideration of all other conditions.

On the other hand, the MLM was less likely to “benefit” non-advanced students who remained tracked for having a similar negative average initial ability due to the initial ability predictor being group-mean centered. In this way, to be below average, the student had to be below his/her class’ average, which consisted of all non-advanced students. Additionally, for this non-advanced group of students, their grouping coefficient was more significant and much more negative. Not only were non-advanced tracked students less likely to obtain a positive initial ability gain predictor, but they also received -9.24 NCE points for being in the non-advanced tracked group. However, due to the conditional nature of an MLM, the -9.24 is not quite what it seems. For example, the two tracked classes had the lowest centered initial class ability averages of -11.28 and -8.93. Recall these values indicate the tracked classes had overall grade six/seven average NCE scores 11.28 and 8.93 NCEs below the overall cohort average. When these grand mean centered class initial ability values combine with the -0.701 class initial ability coefficient, positive predicted gains result. However, these positive contributors do not entirely offset the -9.24 fixed effect for being in the non-advanced tracked group. For example, ignoring other predictors, based on the MLM, the tracked class with the lowest average initial ability receives a positive 7.91 (-0.701*-11.28) NCE gain contribution to condition for initial class ability, which when combined with the -9.24 fixed effect for being a tracked class, yields an overall class ability effect of just -1.33 NCEs. Between the overall descriptive statistics, the matched pairs t-tests, and the MLM coefficients conditioning for various predictors, this study finds significant
Evidence in support of sub-hypothesis number one: non-advanced students perform better in mixed-ability classrooms than in non-advanced tracked classrooms.

**Results Related to Sub-Hypothesis 2**

The study hypothesized that advanced students would perform independently of the type of class to which they were assigned, tracked (7th) or mixed ability (8th). If true, the dependent variable of NCE change score would be negligible, or not statistically different from zero. In this study, all of the advanced students moved from 7th grade high tracked ELA classes to 8th grade mixed ability ELA classes. As such, the only way to interrogate this hypothesis was to perform a matched pairs t test for the average change score difference between advanced students’ six/seven average NCE score and their 8th grade NCE result. Overall, advanced students had an average gain of 1.81 NCE points in the aggregate and 1.50 points for matched differences. While this gain was statistically insignificant, and the 95% confidence interval included the possibility of a small average loss [-0.593, 4.22], the results indicate not only may advanced students’ progress be independent of class assignment, but their scores may also improve when they are learning with a more heterogeneously grouped set of students. The confidence in this possibility is lowered due to the absence of an advanced group of students who stayed in high tracked classrooms against which a contrast could be drawn. Nonetheless, given the results from the two non-advanced sub-populations and this advanced group, it would seem appropriate to infer no harm was done to advanced students in 8th grade, which is a concern raised in the literature. Some researchers have found similar positive or neutral effects of mixing on higher ability students, while other studies have identified somewhat negative effects (see, for example, Argys et al., 1996; Burris et al., 2008; Goldring, 1990).
Absent a control group of advanced students who would have stayed in high tracked classes, another way to evaluate the effect of mixed ability instruction on the advanced students in this study was to consider the regression to the mean phenomenon. In this case, simply due to chance, one would have expected the gains for the advanced students to naturally be below or close to zero. However, as previously indicated, the advanced students had a positive average gain that was almost statistically different from zero. Additionally, the MLM estimated the marginal mean for advanced students moving to mixed ability instruction, while considering and conditioning for all other variables, to be 6.19. Coupled with the actual matched pairs gain, the lack of a detrimental effect on advanced students who are mixed with non-advanced students for instructions seems clear. Indeed, this study provides evidence there may even be an academic benefit for advanced students learning in mixed ability experiences.

**Results Related to Sub-Hypothesis 3**

As discussed in Chapter 3, applying a critical theory lens to the two-year distributions of both race and achievement was another component of the study. Based in the literature, differential effects of grouping students by ability exist and they follow racial and socioeconomic lines in a way that widens already existing gaps. The study hypothesized the racial composition of mixed level classes would better reflect the school’s population than tracked classes.

**Racial Distribution and Types of Classrooms.** Table 4 from Chapter 4 is reproduced on the following page and presents the percent of minority students by class period in both 7th and 8th grades. The 7th grade relationship between class period and percent minority was confirmed at a p-value of 0.000, while the lack of dependency between 8th grade class period and race was also “confirmed” via not rejecting independence, $p = 0.970$. 

149
Table 4, reproduced

Proportion Minority by Class Period, Grades 7 and 8

<table>
<thead>
<tr>
<th>Proportion Minority</th>
<th>Class Period Gr 7</th>
<th>Class Period Gr 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nadv</td>
<td>Adv</td>
<td>Adv</td>
</tr>
<tr>
<td></td>
<td>.79&lt;sub&gt;a,c,d&lt;/sub&gt;</td>
<td>.60&lt;sub&gt;a,b,c,f&lt;/sub&gt;</td>
</tr>
<tr>
<td>Adv</td>
<td>.38&lt;sub&gt;b&lt;/sub&gt;</td>
<td>1.00&lt;sub&gt;c&lt;/sub&gt;</td>
</tr>
<tr>
<td>Nadv</td>
<td>.90&lt;sub&gt;a,c&lt;/sub&gt;</td>
<td>.46&lt;sub&gt;b,d&lt;/sub&gt;</td>
</tr>
<tr>
<td>Adv</td>
<td>.91&lt;sub&gt;c,e&lt;/sub&gt;</td>
<td>.90&lt;sub&gt;c,d&lt;/sub&gt;</td>
</tr>
</tbody>
</table>

1. Tests for differences in proportions are adjusted for all pairwise comparisons within a row using the Bonferroni correction.

Figure 8 presents the overall three-race composition of students in the study, while Figure 9 provides a visual of the incongruence of racial distribution between advanced classes and non-advanced classes in the aggregate.

![Pie chart](image)

**Figure 8.** Overall cohort distribution of African-American, White, and Latino students.
Figure 9. Distribution of African-American, White, and Latino students in advanced and non-advanced 7th grade classes.

Additionally, Figure 10 shows the racial composition for each class period in both 7th and 8th grades. Focusing on the darkest portion of each pie, which represents White students, it is readily apparent how different the 7th grade classes on the left are from the 8th grade classes on the right. Yet the same cohort of students populated both sets of classrooms.

Figure 10. Distribution of White, African-American, and Latino students by 7th and 8th grade class period.
In Figure 11 below, the bar heights show the actual distribution of race in the advanced and non-advanced ELA classes; while the six bar-width dashes indicate what the distribution would look like if race were independent of level. In sum, the findings of this study support prior research conclusions that minority students are over-represented in non-advanced classes and under-represented in advanced classes. Necessarily, the opposite is true for White students. Tracking students by ability has a segregative effect on classroom composition (Ansalone, 2006; Gamoran & Mare, 1989; Oakes & Guton, 1995).

Achievement results linked to race. The dependent variable of interest in this study was the change in students’ standardized test performance, as measured by Normal Curve Equivalent (NCE) points, between their six/seven average and their eighth grade scores. Assessing change scores by race both in the aggregate and as matched differences, respectively, all three racial subgroups showed improvement: White students gained 2.16 and 1.43 ($p = 0.158$) NCE points, African American students gained 3.14 and 2.23 ($p = 0.028$) points, and Latino students gained
4.90 and 3.65 ($p < 0.000$) NCE points. However, only African American and Latino students’ gains were statistically significant, perhaps indicating learning in less racially isolated environments positively affected their outcomes. Figure 12 presents a visual of the previous explanation by showing all groups originating at the same point, time 1 on the x-axis, and the NCE change score results reflected at time 2 on the x-axis, or after 8th grade instruction.

![Figure 12](graph.png)

*Figure 12.* Graph comparing overall NCE change score trajectories by race.

**Results Related to Overarching Research Hypothesis**

Essentially both sub-hypotheses one and two align with the work of Burris et al. (2008), who found lower level students benefited academically from mixed level classes, while advanced students were unaffected by being in mixed ability classes. By combining the discussion around sub-hypotheses one and two, evidence mounts in support of the overarching research hypothesis that how students are assigned to classes in the secondary school setting has a relationship with relative student achievement on standardized state tests. Unlike prior research focusing on how
tracking students by ability affected low track and high track students, this study primarily
examined how mixing students from the low and high tracks together for instruction may relate
to relative student achievement. Whereas prior research found gap-widening effects of tracking,
with low track students being harmed and high track students reaping benefits, this study puts
forth evidence of a different variety. In this case, on average, both non-advanced and advanced
students in mixed ability classrooms both experienced, and were predicted to experience,
academic gains.

See Table 18 for MLM-adjusted marginal means for the three grouping trajectories.
Essentially, this table provides the MLM’s predictions for the average student’s gain in NCEs
from each of the three trajectories. The substantial and statistically significant gains realized by
the non-advanced mixed students hold promise for lessening the pernicious achievement gap.
Importantly, their gains are not juxtaposed with a loss for advanced students, a common citation
for continuing to track students into different classes. Rather, mixed ability instruction appears to
be a neutral practice for advanced students and may in fact have a positive effect. If these
findings were to be replicated and substantiated, the ethical dilemmas of whether to track
students by ability could dissolve.
Table 18

*Hypothetical Marginal Means Based on MLM*

<table>
<thead>
<tr>
<th>Grouping Sequence</th>
<th>Marginal Mean Estimates&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Std. Error</th>
<th>df</th>
<th>95% Confidence Interval</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAT7 stay NAT8</td>
<td>-5.140&lt;sup&gt;b&lt;/sup&gt;</td>
<td>9.023</td>
<td>2.553</td>
<td>-36.916</td>
<td>26.636</td>
<td></td>
</tr>
<tr>
<td>NAT7 to MIX8</td>
<td>5.357&lt;sup&gt;b&lt;/sup&gt;</td>
<td>2.373</td>
<td>3.163</td>
<td>-1.979</td>
<td>12.693</td>
<td></td>
</tr>
<tr>
<td>AVT7 to MIX8</td>
<td>6.188&lt;sup&gt;b&lt;/sup&gt;</td>
<td>2.805</td>
<td>6.059</td>
<td>-.659</td>
<td>13.036</td>
<td></td>
</tr>
</tbody>
</table>

a. Dependent Variable: NCE 8 – 6/7avg.
b. Covariates appearing in the model are evaluated at the following values: Sex = .29, GrdMnCtrClsSz8235 = .6561, Grand Mean Ctr Sex for 8th Class = .0025, Class Centered 67 for 8th = -.5624, Grand Mean Center Class Avg 67 for 8th = -.0411.

Overall, this study found non-advanced students’ one-year gains depended upon the type of class to which they were assigned. Those in the mixed classes had significantly greater gains than students who remained in non-advanced tracked classes. The advanced students, on the other hand, performed in mixed classrooms as they previously had in tracked classrooms. By definition, most students are non-advanced, making the possible impact of de-tracking more substantial. Figure 13 provides a visual of the average NCE gains by grouping trajectory along with the wide gaps between initial NCE values. As evidenced by the stable and similar slopes of the advanced to mixed and non-advanced stay tracked lines, such assignments would maintain the status quo. The steeper and positive slope of the line representing the non-advanced to mixed trajectory points toward the closing of at least some of the achievement gap. Given the significant gains by being in mixed ability classes revealed by this study, along with the plethora of studies indicating the deleterious effects of tracking on non-advanced students, it seems clear how students are assigned to classrooms matters with respect to relative student achievement.
Regarding sub-hypothesis three, this study supports prior research by showing the segregative effect of tracking students by ability.

While admittedly similar to the overarching research hypothesis, another implication of this study is that overall subject area- and teacher-based relative student achievement is related to the grouping structure utilized to create the instructional groups/classes. This distinction is intentionally included to both inform and stimulate dialogue around accountability-based policies that include student achievement metrics to rate teacher effectiveness in high stakes manners. For if classroom assignment practices are related to relative student achievement, then not all of the student gains, or lack thereof, can be directly attributable to any given teacher.

*Figure 13. NCE initial values and change score slopes by grouping sequence.*
Figure 14 displays regression lines by 8th grade class period displaying the relationship between student NCE change scores and their initial ability NCE value. None of the relationships is strong and all have negative correlations, as expected in any regression to the mean scenario. Figure 15 provides the same information without the class regression lines to provide a better view of the points. Two reference lines are also included, the vertical line $X = 54.8$, to show the overall grand mean initial ability, and the horizontal line $Y = 0$ to show expected change scores for students who maintained their prior achievement level NCE. The two lines divide the data points into four quadrants. The vast majority of the points are above the zero reference line, indicating positive change scores, or gains. Notably, the quadrant with the fewest points is the lower left, which represents students who started 8th grade below average and had negative change scores at the end of the year. Equally important and related to the prior statement, the upper left quadrant contains the most data points, all representing students who also began the year below average but had positive change scores at the end of the year. The other half of the data points are more evenly distributed around the zero change score line, though a greater number of the students who began the year with above average ability also ended the year with positive gain scores, located in the upper right quadrant.
Figure 14. NCE change scores regressed on prior ability proxy across 8th grade classes.
Figure 15. NCE change scores versus prior ability proxy across 8th grade classes with reference lines.

Implications of Results

An aim of any research is to stimulate reflection and dialogue, at a minimum, or lead to policy changes and future research, at a maximum. Within the limitations of this research and under similar conditions, this study’s results, along with a broad base of prior research, point to a need for specific kinds of change within the American public education system.

Implications for School Districts

Within the limitations of this study, the results indicate school districts may need to review their approach to how secondary students are assigned to classrooms for instruction. If district leaders are not currently asking questions of their principals regarding their class
assignment practices and resultant racial and achievement distributions, they should begin doing so. As applied in this study, critical theory needs to be alive and well within pre-K-20 education, particularly when the outcomes have been persistently disparate for both minority and low-income students. The research overwhelmingly indicates when students are tracked by ability for instruction, the achievement gap between both minority and low-income students and their White, higher income peers widens (Lee & Bryk, 1989; Oakes & Guiton, 1995). It has become far too easy to shirk responsibility for discrepant results, pointing to outside of school factors. While schools cannot address every challenge or fix every problem arising from societal constructs, they can undoubtedly work to ameliorate the impact of these challenges. Educational structures and systems are notoriously change-resistant (Cuban, 2004; Lortie, 1975; Ogawa, 2009; Wagner, 2008). Policies and practices originating in the post-Civil War era, such as tracking, likely no longer serve the purported greater good of public education.

Policy Considerations. Applying an interrogative critical theory lens toward all policies guiding a school district should be part of the extant culture. However, with respect to grouping students by ability for instruction, there is evidence such questioning is not occurring. The majority of secondary schools continue to track students by ability, even if only in a course-by-course manner. Yet the preponderance of evidence indicates the practice is not only ineffective but also harmful to the most vulnerable students (Archbald et al., 2009; Kozol, 2006; Loveless, 2013; Oakes, Gamoran, & Page, 1992). The maintenance of such status quo would not be accepted by a culture of critical theory. It is incumbent on district leaders to disrupt the status quo in favor of innovation and equity.

To specifically begin to address policies regarding how students are assigned to classrooms, district leaders could first request data regarding classroom composition with respect
to both race and achievement outcomes. Such data could be used to begin reflection and conversation around policies leading to classroom distributions of both achievement and race. Given the literature around differential expectations and instructional strategies based on level of class taught, it would also be instructive to facilitate some kind of audit to gather relevant data and information in this regard as well. Policy changes or alterations are best constructed with a full understanding of the current reality. Working from this information, conversations could begin and additional study of relevant research could be provided. A starting point to a more mixed level approach to assigning students to classes could be to blur the lines of assignment cut-off points along with examining how and why borderline assignment decisions are made. Ultimately, what is most critical for all students is the teacher in the classroom providing quality, aligned instruction utilizing various pedagogical approaches.

**Practice.** Importantly, changes in policy do not always directly translate into changes in practice. A troubling example of such a disconnect is how tracking became more widely applied following the *Brown v. Board of Education* (1954) ruling that ordered schools to no longer segregate students (Chayt, 2010). Tracking and, more recently, charter schools and vouchers, have been viewed as tacit tools to maintain segregation. Yet mixing students by race and ability will not automatically translate into more equitable outcomes. Attention must be given to instructional practices that better meet varied student needs. Substantive quality staff development must be provided to teachers such that they are able to be effective instructors for all students. Indeed, research has indicated, including within this study, that quality instruction can overcome most any situation, including a less than optimal classroom assignment design (see, for example, Haycock, 1999). Professional development alone will also not suffice. Administrators must be instructional leaders and support the implementation of altered practices,
encouraging collaboration among teachers. As evidenced in Figure 16, variation in student gains across classes clearly exists, though not in significant ways, likely due to the single cohort nature of the study. The two tracked classrooms are to the left and the six mixed classes are represented with two bars each, the darker one representing the non-advanced to mixed students. In every case, regardless of the time of day or teacher, the non-advanced students outgained their advanced peers. If alterations in student to classroom assignment policies to include more mixing of secondary students for instruction would yield similar results across all classrooms, the ever-present achievement gap may begin to diminish. At the same time, by improving upon teachers’ differentiation techniques, they would be able to better meet the needs of students regardless of their level or the grouping structure employed, increasing gains universally.

![Figure 16. 8th grade NCE gains by class period and grouping sequence.](image-url)
Implications for Universities and Other Preparers of School Leaders

The results of this study bring implications for those who prepare school and district leaders. In broad ways, these implications mirror those for school districts themselves. A discussion of these implications and considerations follows.

Research. Higher education has the opportunity to play a unique role in bridging the gap between theory and practice. Universities have a plethora of research they could share with district officials who typically do not have the time needed to read the literature let alone distill it into actionable knowledge. Universities can serve districts by instituting collaborative efforts to present critical and targeted findings to district leaders. In this case, an overview of the literature regarding tracking and its relationship with the achievement gap would be of primary interest. Such exchanges could be accompanied by recommendations for examining current practices with an eye toward leading principals and teachers to possible alterations in practice.

Leadership preparation. While preparing future educational leaders, a large focus should be given to developing their ability to lead both instructionally and culturally. Particularly given the readily available and disaggregated data stemming from NCLB, the impetus for change is stark. Future leaders must be able to envision different outcomes before they can lead others to them. Reviewing the research around classroom assignment practices while also acknowledging principals’ and district leaders’ power to alter such practices could play a critical role in the future of education. At a minimum, focusing on quality instruction for all students is paramount. Future leaders must not only be well versed in research-based instructional practices but also equipped to lead and support teachers enacting them.

While not the focus of this study, one clear indicator of how critical a teacher is to a student’s outcome was the consistent statistical significance of teacher sequence on student NCE
change score. Perhaps even more telling, as seen in Table 19, two of the most statistically significant teacher sequence pairwise comparisons were that of sequence AC compared to BC \((p = 0.007)\) and AD compared to BD \((p = 0.075)\). In both comparisons, the 8\(^{th}\) grade teacher, C in the first pair and D in the second pair, were held constant, yet students had statistically different outcomes with the same 8\(^{th}\) grade teacher, seemingly attributable to their 7\(^{th}\) grade teacher, either A or B. Remarkably, then, students’ 7\(^{th}\) grade teachers had a greater affect on their 8\(^{th}\) grade NCE change score than did their 8\(^{th}\) grade teacher. This result underscores the importance of having a quality teacher, as their effects may linger into the future. To be more certain of the latent effect of prior teacher, an investigation into whether students tended to have gains or losses with teachers A and B is warranted, as there is a tendency for losses one year to be countered with substantive gains the next year (Rothstein, 2009). In either case, teacher effect is clearly significant.
Table 19

MLM-based Teacher Sequence Pairwise Comparisons

<table>
<thead>
<tr>
<th>Pairwise Comparisons*</th>
<th>(I) Teacher Sequence</th>
<th>(J) Teacher Sequence</th>
<th>Mean Difference (I-J)</th>
<th>Std. Error</th>
<th>df</th>
<th>p.c</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>AD</td>
<td>5.683</td>
<td>3.156</td>
<td>3.585</td>
<td>.154</td>
<td></td>
</tr>
<tr>
<td>BC</td>
<td>AD</td>
<td>6.170</td>
<td>2.249</td>
<td>160.000</td>
<td>.007</td>
<td></td>
</tr>
<tr>
<td>BD</td>
<td>AD</td>
<td>9.343</td>
<td>2.959</td>
<td>2.790</td>
<td>.056</td>
<td></td>
</tr>
<tr>
<td>AD</td>
<td>AC</td>
<td>-5.683</td>
<td>3.156</td>
<td>3.585</td>
<td>.154</td>
<td></td>
</tr>
<tr>
<td>BC</td>
<td>AD</td>
<td>.487</td>
<td>3.339</td>
<td>4.542</td>
<td>.890</td>
<td></td>
</tr>
<tr>
<td>BD</td>
<td>AD</td>
<td>3.660</td>
<td>2.045</td>
<td>159.160</td>
<td>.075</td>
<td></td>
</tr>
<tr>
<td>BC</td>
<td>AC</td>
<td>-6.170</td>
<td>2.249</td>
<td>160.000</td>
<td>.007</td>
<td></td>
</tr>
<tr>
<td>AD</td>
<td>BC</td>
<td>-5.683</td>
<td>3.156</td>
<td>3.585</td>
<td>.154</td>
<td></td>
</tr>
<tr>
<td>BD</td>
<td>BC</td>
<td>3.660</td>
<td>2.045</td>
<td>159.160</td>
<td>.075</td>
<td></td>
</tr>
<tr>
<td>BD</td>
<td>AC</td>
<td>-9.343</td>
<td>2.959</td>
<td>2.790</td>
<td>.056</td>
<td></td>
</tr>
<tr>
<td>AD</td>
<td>BC</td>
<td>-3.174</td>
<td>3.127</td>
<td>3.434</td>
<td>.376</td>
<td></td>
</tr>
<tr>
<td>BD</td>
<td>BC</td>
<td>-3.174</td>
<td>3.127</td>
<td>3.434</td>
<td>.376</td>
<td></td>
</tr>
</tbody>
</table>

Based on estimated marginal means*. The mean difference is significant at the .05 level.
a. Dependent Variable: NCE 8 – 6/7avg.
c. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

The results of this study also highlight the need for strong cultural leaders in public schools. For instance, from this study, given the wide variation in student outcomes whether students were tracked by ability for instruction or in mixed level classes indicates factors other than student ability at play when determining tracked class assignments. Based on the literature, one may hypothesize cultural constructs are contributing to such decisions, and those constructs tend to favor White middle and upper class students being assigned to higher level tracks irrespective of standardized test performance. For instance, students may be recommended for an upper level class based on their solid work ethic or positive classroom behavior, namely being a “good student.” Such definitions derive from the hegemonic culture and so reproduce the familiar status quo. To effectively disrupt such tendencies without alienation, leaders must be
astutely aware of the extant culture of their school and be able to shepherd toward change (Darling-Hammond, 2010).

There are wide held beliefs that teachers can be more effective and efficient in their practice when they are teaching students more homogenous in ability, a purported outcome of tracking (Ansalone, 2009; Ansalone & Biafora, 2010; Argys et al., 1996; Gamoran, 2009). However, the achievement and ability in any given class is likely more variable than thought. In this study, the variation in student achievement on the state standardized test varied substantially within classrooms at both grade levels. In that 7th grade tracked students by ability, one would have expected considerably lower classroom level variation than in 8th grade mixed ability classrooms. However, the average classroom standard deviation for 7th grade scores was 14.5. While smaller than the average variation in 8th grade classrooms, which was 17.7, the difference was not sizeable. Additionally, in 7th grade the three advanced sections were more discriminating with regard to perceived student ability, yet two of the three advanced classrooms had higher than average outcome variation. In 8th grade, when six of eight sections were fully mixed with non-advanced and advanced students while two remained solely comprised of non-advanced students, one would have expected the non-advanced tracked classes to be less variable in outcome than their mixed counterparts. However, each of the low and high extremes in classroom standard deviation of achievement, 14.1 and 19.7, represented the tracked classrooms. The educational institution is therefore holding on to a past practice, tracking, that may not even be serving its purported goal, providing for more targeted instruction.

Limitations of the Study and Recommendations for Future Research

This study provided a thorough analysis of a unique case in which many middle grades students experienced both tracking and mixing for ELA instruction in sequential years while
attending the same school. However, the study was also limited in several ways, primarily in relation to the sample and methodology. These limitations will be presented along with recommendations for future research, which include ways to capitalize further on data from the current study’s site as well as ideas to extend the interests of the study to other contexts.

Limitations

Sample. A number of issues could be improved upon regarding the sample for this study. As has already been discussed, studying a magnet school raises generalizability concerns due to the possible lack of representation in the sample. Additionally, only approximately 200 students’ results were in the study, which represented just one cohort of students. While no known disruptive events occurred over the course of time from which student outcomes were collected, there may be something particularly unique about this span of years that remained uncaptured via the analytics employed. By having additional student data in the study, the results would be more reliable. In part, the study only examined one cohort due to the researcher’s familiarity with time-varying school level matters. This familiarity served as both an asset to the study and a limitation. One asset in this case was knowledge that the particular cohort studied experienced a stable school environment over the course of data production. All teachers in the relevant grades had at least several years of experience, did not have student teachers, and had no irregularities in attendance. In adjacent cohort years, such stability was not present, which impacted the decision to only study the single cohort.

Methodology. The first methodological limitation of this study relates back to the sample itself, specifically the sample size. The full utility provided by multilevel models (MLM) was not realized in this study, largely due to only having eight units, or classrooms, at level two. Some experts suggest a minimum of ten level-two units for the MLM to best function (see, for
example, Snijders & Bosker, 1994). As such, the MLM yielded coefficients to appropriately condition for potentially confounding variables but it was unable to identify enough variation to allow for random effects beyond the 8th grade classroom nesting structure. In other words, if there had been a larger sample at level two, meaning more classrooms and more between group variation (BG), it may have been possible to identify and, importantly, explain other random effects. The model could then answer such questions as does the effect of initial student ability vary across classrooms? By being able to include such elements, the model would have been able to provide a better overall fit to the data and to answer additional questions of interest.

Furthermore, the model was also not able to determine any cross-level interactions. An example of a potentially interesting cross-level interaction in this study would be whether an individual student’s sex played a role in their grouping trajectory effect. It is possible some of these cross-level interactions, or interactions at all, existed but, again, the variation did not substantiate their estimation.

The study could have been made more robust by including additional quantitative metrics in addition to the annual North Carolina End of Grade (EOG) test results scaled into NCE units. There were clear benefits to using the EOG metrics in that they resulted in a type of census as the tests were required to be taken by all students. Additionally, as discussed in Chapter 3, the assessments undergo a rigorous development process and are under continual review, subject to statistical testing and equating. Furthermore, the tests are high stakes, administered under controlled environments, and both teachers and students take them seriously. All the same, a primary concern regarding the sole use of the EOG is that it is only taken one time at the end of a year of instruction. Numerous studies indicate summer learning loss is a real issue, particularly among lower achieving and lower socioeconomic students (see, for example, Cooper, Nye,
Charlton, Lindsay, & Greathouse, 1996). If a group of students systematically differs in their learning progress or loss during shared outside of school time, such as summer, it would seem using a year-end metric to measure growth over the next year is limited on its face. Perhaps better indicators of change due to the current year’s instructional experience would be obtained by administering a beginning of year assessment, with possible benchmarks throughout the year, then culminating in an end of year assessment similar in scope and scale to the others. Such a study could extend this study’s work and may offer more accurate insight into the variables of interest.

By only including quantitative variables, the study was limited in scope, particularly considering the sociocultural belief systems undergirding classroom assignment practices. Including qualitative aspects in the study regarding teacher and student meaning-making and perceptions of the teaching-learning experience in the different grouping environments would contextualize the study. By working to uncover the layered realities of the educational experience as it relates to grouping practices, the study would have the potential to have an emancipatory effect. Semi-structured interviews and observation field notes could work together with the quantitative data to bring voice to all parties, particularly those historically oppressed (Marshall & Rossman, 2011). Furthermore, by valuing all stakeholders’ opinions, experiences, and input, “the critical ideology needed to offer possibilities of hope to students and communities who have been marginalized or otherwise silenced by our system of schooling in the United States,” can develop (Dillard, 1995, p.560).

Knowing the complexities behind the teaching and learning process, the prevalence of tracking, and the socially reproductive nature of schools, it would be assistive to also seek to understand student and teacher experiences in tracked and mixed ability classrooms. By linking
the constructed realities to the quantitative data, a more complete picture of the reality of schooling practices may develop, from which possible policy implications could arise. Looking for the convergence of qualitative and quantitative data around the same topic offers a way to triangulate, or lend more credence to, the results. Furthermore, by simultaneously utilizing the best aspects of qualitative and quantitative methodologies, some of the challenges each method presents could be counterbalanced by the other method’s strengths (Jick, 1979). Therefore, including such qualitative aspects into the design could mitigate the limitations of employing a quantitative-only study.

**Recommendations for Future Research**

**Further Exploit Unique Site.** Situating the study in a magnet school may lessen the generalizability of the results in part due to selection bias. However, maintaining the comparisons within the school and largely within each student, meaning students effectively served as their own control, effectively mitigated much of that bias. The unique opportunity to compare students’ achievement relative to themselves and to different treatments outweighed concerns regarding studying within a magnet school. Nonetheless, there may be something systematically different about the students who won the entrance lottery and enrolled at the school. Not only did students have to enroll, but to be included in this study, they had to be at the school at least for both seventh and eighth grades. Following the work of Ballou, Goldring, and Liu (2006), an improvement to this study would be to obtain information about students who won the entrance lottery but chose not to attend the school or who attended but then left the school. Following and comparing these two groups of students with the ones in the study would add an interesting contrast regarding differences between families/students who apply for
entrance to this school but do not end up attending either by not winning a spot, turning down a spot, or attending and then leaving the school.

Should 6th grade ELA and 9th grade English outcome data from this study’s same students be available, an interrupted time series model analysis would offer another perspective on the relationship between student to assignment practices and relative student achievement. Again, data from this site offers a unique situation in that a) most students attend the school from 6th through 12th grades, and, b) for ELA in 6th and 7th grades, students are fully tracked by ability, in 8th grade all but two of the eight sections are mixed ability, and in 9th grade all students are assigned to mixed level English classes. In this case, by employing an interrupted time series with four data points for each student, time invariant predictors could be considered accounted for within subjects and the effect of treatment could be observed as a trajectory change between grades seven and eight (Raudenbush, 2009). With data from grades six through nine, similar to this study, matched pairs analyses could also be employed wherein the differences would come from subtracting the grade six/seven average from the eight/nine average NCE scores. Using averages instead of single year results, as was the case with the grade eight results in this study, could ameliorate both some of the natural variability within individuals over time as well as the peer and teacher effects between groups. In this proposed analysis, the grade six/seven average would represent student performance while tracked for instruction, while the grade eight/nine average, for almost all students, would represent student performance while in mixed classes.

Similar with regard to data inclusion but different with respect to technique, another way to capitalize on this unique site would be to utilize a repeated measures design. Such a method was initially considered for this study. However, by using essentially two values in a repeated measures design, many of the benefits of repeated measures would be lost, as variation between
two points may simply be random rather than indicative of a variable’s effect. For this reason, the two-level MLM was utilized (Hoffman, 2015). Nonetheless, with data for the same cohort of students traversing from grade six through eleven at which point all students take the ACT, a multilevel model design with repeated measures at level one, students at level two, and classrooms at level three could offer additional insight into relationships between and within classroom assignment practices. At this same secondary school, beginning in eleventh grade, students can opt to take Advanced Placement English classes, which effectively siphons off a similar group of advanced students, as in grades six and seven, from the remaining English classes.

Another possible extension with this same data set would be to employ an adaptive centering model with random effects, as described by Raudenbush (2009). Such a model is a kind of value-added model and would control for both unobserved student and school or classroom level confounding variables. By using a value-added model (VAM) similar to those used to evaluate teacher and school effectiveness, significant accountability policy implications could arise. There is a burgeoning body of literature around VAMs, including their use to model student achievement gains and their legitimacy regarding the attribution of those gains to teachers and administrators. However, few studies are employing VAMs to investigate possible relationships between structural elements, such as student assignment practices, and VAM results. Such studies are critically needed, particularly in light of the high stakes accountability decisions being made based on VAMs.

Two other follow-up studies would further contextualize the results. First, given the site of the study was a magnet school, it would be interesting to include a student’s zip code as a possible predictive factor for student achievement. Housing policies are one of the remaining
vestiges of segregation. While zip codes are quite broad indicators of housing, it would be instructive to know if geographic location offered information about student performance. Additionally, in that this study indicated advanced students may also benefit from being in mixed ability classes, which was counter to prior research, additional investigation into advanced students’ performance and characteristics is warranted. Perhaps what appears to be the over-identification of advanced students within this district plays a role in the results. If, as other studies have indicated, truly gifted students need separate learning environments, then a finer examination of the top performers’ NCE change scores would be informative. Investigating both how and why students were assigned to advanced classes in the first place would also offer insight into the constructed culture of both the school and the community.

Specifically related to this study and similar to the work of Rothstein (2009), it would be interesting to re-analyze this study’s data in a reverse-time manner. In other words, what would the MLM show if the dependent variable of interest had been the change between students’ 7th grade outcome, when all students were tracked, and their average outcome from grades eight and nine, when most students had been mixed for two years of ELA instruction? Perhaps a similar fallacy would be identified, such as students’ ninth grade class period, when all students were mixed, being predictive of their retroactive change score. Finally, another interesting analytic option would be to use randomization and bootstrapping methodologies to estimate effect sizes for the various predictors involved in this study (Boucher et al., 2010; Hox, 2010). Randomization techniques take the observed data values and, without replacement, randomly assign each to one of the treatment groups, whereas bootstrapping involves treating the observed data as the population and samples from it with replacement to create a new data set. Both techniques allow for estimation of how likely the observed set of data would be by chance alone,
the primary goal of statistical inference. As software and technologies have advanced, the practicality of utilizing such mass simulation techniques has become both manageable and desirable to many statisticians (Berry, Johnston, & Mielke, 2014).

**Examine Teacher Combined Effects.** An under-investigated area of research is the potential effect of combinations of teachers on an individual teacher’s achievement effects. For instance, in this study, are the 7th grade ELA results, after conditioning for other factors, attributable to the ELA teachers or to their team of teachers? For example, it would seem a humanities teacher’s instruction and assignments related to reading and writing could play a role in students’ ELA test scores, which are largely reading comprehension examinations. If true, another set of implications arises regarding using student achievement results to evaluate teachers. If it is not possible to isolate the contributions of one teacher, then it is not fair or equitable to proceed with achievement-based accountability.

**Extend Study Beyond Current Site.** There is a paucity of literature examining how course to course tracking, also referred to as neotracking, may relate to student achievement (Lucas & Berends, 2002; Mickelson & Everett, 2008). While tracking remains commonplace in secondary schools, students are less likely to be compartmentalized all day, which may mediate some of the effects reported by prior research. Therefore, a new batch of studies is needed to comprehensively examine the similarities and differences between student outcomes from courses in which students are tracked and courses in which they are mixed by ability for instruction. In a similar vein, another interesting study would be to examine the achievement outcome variation in tracked and mixed ability classrooms. If, as was evident in this study, the range of abilities in tracked versus mixed classrooms is insignificant, what possible goal could be achieved by sorting students?
As discussed previously, it would be beneficial to extend this study beyond the current magnet school site to include both non-magnet and geographically diverse schools. If this study’s findings were corroborated, implications for policy and practice would be more solidified and actionable. If discrepant results were observed, then attention could be directed toward distinguishing variables related to positive, neutral, or negative relationships between tracking and relative student achievement. In either case, having a larger base from which to analyze results and draw inferences would enhance the study and make results more applicable to additional contexts.

Additionally, this study examined how students previously tracked by ability would respond performance-wise when placed in mixed ability classrooms. Two other types of studies would help to round out the implications for practice. The first type would examine how students who had always been mixed for instruction responded when tracked by ability the first time. In addition to analyzing subsequent results, it would be interesting to see if initial ability discrepancies may be smaller than those observed in this and other studies. Such a finding would be parallel to the work of Gamoran and Mare (1989) who found the longer students are separated by ability for instruction the more disparate the outcomes. The second proposed type of study would examine the distribution of achievement by race and initial ability for students who spend their whole educational experience in mixed ability classrooms. In such cases, do the intractable racial and socioeconomic achievement gaps exist, and, if so, to the same degree?

**Conclusion**

In conclusion, this study sought to investigate possible relationships between secondary school student-to-classroom assignment practices and relative student achievement on standardized tests in addition to residual effects of such practices on classroom racial
composition. The hypotheses were conceptualized by both social learning theory (Bandura, 1977; Vygotsky, 1978) and critical theory, as commonplace practices of student assignment were interrogated. Social learning theory tenets drove the overall hypothesis that there would be a relationship between how students were assigned to classes and their relative achievement. To summarize, the following points were supported by matched pairs inferential statistics, which capitalized on the study’s unique opportunity to follow the same students through both tracked and mixed-ability classrooms, as well as by advanced statistical multilevel modeling to account for inherent nesting structures and to condition for other effects:

- Non-advanced students performed better on ELA standardized tests after instruction in mixed ability classrooms than after instruction in tracked classrooms.
- Advanced students performed similarly on ELA standardized tests regardless of the composition of the classroom.
- Minorities were disproportionately enrolled in non-advanced ELA tracked classrooms, whereas mixed ability ELA classrooms mirrored the overall school racial distribution.
- Latino and African American ELA students performed better on ELA standardized tests after instruction in mixed ability classrooms than after instruction in tracked classrooms.

The lack of racial representation across the advanced and non-advanced students is undeniable. While unknown if the significant gains realized by non-advanced students in mixed ability ELA classes would persist and, importantly, if similar additional gains would occur with continued mixing, this study offers hope for the diminishment of the achievement gap, and, subsequently, the socioeconomic divide. Nonetheless, the research is clear that tracking by ability has multiple negative effects on lower-tracked students (Carbonaro, 2005; Lleras & Rangel, 2009; Welner & Oakes, 1996; William & Bartholomew, 2004). Fewer studies have investigated the effects of
mixed level classes on student achievement, particularly in the middle grades. This study begins to fill that gap in the literature and illuminates the possibilities afforded by altering centuries-old practices around how students are grouped for instruction.

Additional related research is needed to get a better grasp on how both advanced and non-advanced students respond to mixed ability instruction. The vast majority of secondary schools track students by ability making such studies harder to undertake, yet the need is great. The segregative impact of tracking and the institutionalization of racism sit at the center of the arguments for and against tracking (Ladson-Billings & Tate, 1995). Decisions regarding classroom assignment practices may have far-reaching consequences ranging from daily school performance and experience to annual income as an adult. As such, great care and consideration must be given to how and why such decisions are being made, and a process of continuous data review should be instituted.

Some posit the reason tracking nonetheless persists is due to parents of the advanced students pressuring school officials to keep things as they were when they attended school, separated. Such parents possess the cultural and social capital to influence administrators and are largely opposed to detracking (Ansalone & Biafora, 2010). Indeed, to move forward, the meritocratic power base must shift but will not do so without explicit and steadfast leadership of equity-minded school officials. Society can no longer afford to perpetuate class division and social reproduction by way of its institutions (Darling-Hammond, 2010). NCLB has exposed the disparate achievement outcomes occurring en masse in American public schools. The perpetual economic crisis, slippage in international rankings, and fear of other nations’ obtaining world power status together present a unique window of opportunity for both study and change. Due to the nature of the practice of tracking and its connection with race, sitting at the core of people’s
perspectives is a belief system around the ideas of intelligence, culture, ability, and motivation and whether or not such assets are viewed as fixed or malleable. Yet, as cautioned by Ansalone (2003), Hochschild (1997), and others, detracking is not and will never be a panacea for myriad challenging issues plaguing schools and society today. Without great care and intention, administrators leading detracking efforts could see good intentions go awry. As Rubin and Noguera (2004) asserted, centering educational change in a message of improved opportunities and achievement for all students while simultaneously training and supporting both teachers and parents, some of the entrenched practices such as tracking could be left behind.

There is little to no evidence to suggest a movement away from increased accountability with regard to education. The testing industry has experienced a boon, particularly since the advent of required rather than suggested testing arising from NCLB. Given the undeniable influence of lobbyists and partnerships, test-makers’ vested interest in the continuance of mandated testing will likely play a role in future decisions. With the advent of quasi-national standards and forthcoming national assessments, the alteration of such accountability seems even less likely. Hopefully, without the ability to escape from the resulting data, the overwhelmingly inequitable outcomes will eventually lead to progressive alterations in practice that may better serve the public good.

This study opened with the overarching question: “Does how students are assigned to classrooms matter?” Unequivocally, in that upwards of 90% of secondary schools track students by ability, the guiding belief is that yes, it does matter (Archbald et al., 2009). However, this study and myriad others indicate the manifestation of this belief is directionally misguided. That is, it does matter how students are assigned to classrooms and evidence suggests the most equitable way to do so involves some level of mixed ability grouping, the opposite of what is
largely transpiring in secondary schools today. This study will close drawing on the wisdom of Albert Einstein, often credited with the following definition: “Insanity is doing the same thing over and over again and expecting different results.” One could argue continuing to separate students by ability for instruction and expecting anything other than the continuation and expansion of the achievement gap along with income inequality is insane. It is time to trouble the status quo for the benefit of all. Exceptional leaders must be prepared, then supported, to address the social and cultural shifts incumbent upon them to enact.
APPENDIX: NORMAL CURVE EQUIVALENTS

Source: https://ncdpi.sas.com/
References


Berkelman v. San Francisco Unified School District, 501 F.2d 1264 (9th Cir. 1974).


North Carolina Department of Public Instruction. (1999). ABCs announcement materials for local school systems. Raleigh, NC: Division of Communications and Information.


Plessy v. Ferguson, 41 L. Ed. 256 (1896).


U.S. Const, amend. X.
U.S. Const, amend. XIV, § 1.


