STANDARDIZED DISTRACTION:
WHY THE EMPHASIS ON HIGH-STAKES TESTING CAN'T RESOLVE EDUCATIONAL INEQUALITY

by

Jason Giersch

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Public Policy

Charlotte

2012

Approved by:

____
Dr. Roslyn A. Mickelson

____
Dr. Suzanne Leland

____
Dr. James Lyons

____
Dr. Stephanie Moller Smith

____
Dr. Elizabeth Stearns

ABSTRACT

JASON GIERSCH. Standardized distraction: Why the emphasis on high-stakes testing can't resolve educational inequality. (Under the direction of DR. ROSLYN A. MICKELSON)

Standardized testing has exploded into nearly every grade level and subject area in public education over the past two decades. Attaching high stakes to standardized test results was intended to improve education for all students, but especially for those who belong to consistently low-performing groups. Theory and experience, however, show that privileged groups maintain their advantages even in the face of education reforms. Tracking practices in particular have the potential to worsen the inequalities associated with high-stakes testing. This study uses a unique longitudinal dataset to observe the existence, growth, and harm of achievement gaps through high-stakes testing in North Carolina. The study demonstrates that high-stakes standardized tests predict college performance for students whose high school experiences were in in the top academic tracks, but not for students in lower academic tracks. Findings suggest that standardized test scores of lower track students reflect less learning of the kinds of substance and higher order thinking skills needed to excel in college.

DEDICATION

To Tessa and Colby

ACKNOWLEDGEMENT

A number of people helped with this project, but without the patience, encouragement, and sacrifice on the part of Beth Whitaker its completion would not have been possible.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1: INTRODUCTION

The era of No Child Left Behind (NCLB) is an exciting time for research on education policy. Never before has so much empirical data been available on so many aspects of students and schools. Of course, schools don't collect that data for the sake of policy analysts. They do it because federal and state laws require schools to test virtually every student on their rosters, report the results, and then receive consequences, be they commendations and bonuses or warnings and sanctions. Proponents of NCLB and related state laws argue that high-stakes tests give students, teachers, schools, and districts incentive to work harder than they otherwise would. Opponents warn that high stakes tests create incentives and pressures that distort measures of achievement and cause harm to the educational experience of many students.

More than a decade after the passage of NCLB and a quarter-century after *A Nation at Risk*, the federal report that drew widespread political attention to the issue of achievement in public schools, the deadline of 100% student proficiency looms near and progress remains mixed. Stakeholders complain of damaged teacher and student morale, cheating scandals, and the financial cost of frequent assessments, even though some schools, districts, and states claim improved scores. NCLB, in its current state, has lost popularity, but the desire to hold educators accountable through the standardized testing of students continues.

Policies, naturally, do not take effect in a vacuum. Public education in the United States has a long history, firmly-established institutions, entrenched interest groups, and well-established traditions. Our systems of public schooling are complex, and the advent of the latest form of accountability in the form of standardized tests and high-stakes consequences has the potential to correct some problems but also exacerbate others. Perhaps the most-often cited complaint about unintended consequences of accountability policies is the ways in which schools that are already disadvantaged, whether by few resources, unsupportive communities, poor leadership, low achievement, or the like, are forced to compete with schools that have far more advantages (J. Lee & Wong, 2004). Critics of the law have also addressed the adverse impact on vulnerable populations, in particular the poor and racial minorities, who are less likely to have resources in the school, community, or home to give them a fair shot at competing with wealthy or white students (Madaus & Clarke, 2001). These inequalities between schools and between subgroups have been thoroughly documented and continue to receive attention from researchers and policymakers.

The vast differences in student backgrounds make it difficult to hold all students to a common standard of achievement, and the fact that poverty tends to be concentrated in particular neighborhoods and schools makes a fair system of evaluation that much more difficult. Jonathan Kozol’s (1992) book *Savage Inequalities* gave readers a tour through a series of public schools that varied widely in terms of their resources, leadership, faculty, and student bodies. Where I grew up in the suburbs of Philadelphia, North Penn High School boasts two auditoriums, three swimming pools, multiple sports fields, an entire wing devoted solely to fine art, and a planetarium. High salaries attract large pools of applicants and the active parents proudly support championship-winning sports teams and music programs. North Penn scores high on the state’s standardized tests. Just a few miles south, Norristown High School struggles with a tight budget and a poor reputation. Few middle or upper class families still live in Norristown, even though it is the county seat. Its position along a river and light rail line has failed to attract very many commercial or residential developers, save for bail bond shops and subsidized rental property speculators. Test scores consistently trail those in the surrounding school districts. A student who enters the halls of Norristown likely already faces a lot of obstacles in life, and for the four years or so of high school he or she will likely sit among other students who have had similar disadvantages.

There is another source of inequality, however, that exists within individual schools and within social classes and races: academic tracking. The segregation of students within a school into different classrooms according to ostensible academic ability, promise, or interest is a common and popular school practice that is rarely questioned, even though researchers have pointed out that the practice often separates students along lines of race and class and results in differing opportunities to learn (Oakes, 1986). Studies have shown that while academic records do play a role in the sorting of students into different tracks, other characteristics of the students also determine academic track (Gamoran, Nystrand, Berends, & LePore, 1995). Students in upper-track classes are more likely to receive assignments that are more engaging, open-ended, and challenging. Students in lower-track classes are more likely to be receive lessons that require little, if any, higher-order cognitive skills like analysis, synthesis, or evaluation. Instead, lower-track students receive lessons that involve rote learning, simple recall or recognition, and the mere parroting of the instructor or instructional material (Page, 1991; Watanabe, 2008).

Consider again the examples of North Penn and Norristown. Setting aside the impressive facilities of the wealthy district and the concentrated poverty in the other, when students take honors classes in either school, they may find themselves exposed to resources, lessons, and peers that differ markedly from the non-honors classes down the hall. When they enroll in an Advanced Placement (AP) course at either school, they likely encounter a level of expectations that would rarely be found in another level of the same course. Given the attention that inequalities between schools receive, it is remarkable that the inequalities between classes within schools get so much less consideration.

My own teaching career has spanned a wealthy suburban school district, a selective private school, a struggling charter school, and a high-minority urban public school. In each of these settings I have taught both honors and regular classes. In most cases I enjoyed the flexibility of creating my own lessons and assessments. The greatest exception came when I taught Civics and Economics, a tenth-grade course that was part of North Carolina’s high-stakes testing accountability program. For three consecutive years I taught the course to both honors and regular students. Most days I used exactly the same lesson plan for both tracks, from the anticipatory set to the independent practice. Usually the lessons would work equally well in the two classes, such as when I had students re-create a passage of their choice from the Declaration of Independence in some creative way. Other times they would succeed in one and fail in the other, such as when I read Dr. Seuss’s *The Lorax* and asked them to identify the different means of production found in the story. Honors kids were willing to be read to; the regular classes were not interested.

All the while, the messages from my administration were clear. My state-tested classes were the most important and the others did not matter. I should push the honors students to score well but more important was to make sure that all students at least made proficiency. My colleagues and I studied the old exams, reviewed the standards, and shared “effective” lessons. A lesson was considered especially valuable if (1) it covered content that was likely to be found on the exam, (2) it required the student to write something about that content that could then be stored in a portfolio,[1] and (3) it was sufficiently brief and interactive to keep the attention of even the most jaded, disaffected, or low-skilled student. Over time we developed what we called the “interactive notebook”, which was really just a series of standards-based worksheets that students at any level could complete. We would present a new one to our students each day. If the honors students finished in 15 minutes, the rest of the time could be used for debates, mock trials, essays, and games. In our regular classes, we were usually content to allow the students to spend the entire class on the worksheet.

Accountability policies, such as those using high-stakes tests, are intended to boost achievement across all schools and all classrooms. Whether placed at a North Penn or a Norristown, in an honors class or regular class, the assumption is that students, teachers, and administrators could improve their productivity if their progress was measured, analyzed, and reported. As a consideration for the differences among student backgrounds and school resources, accountability schemes rate schools by their “growth” and by their “progress.” Such calculations do little to ease the stress that test takers and

[1] In North Carolina’s high-stakes accountability program, students who repeatedly fail a standardized test can sometimes submit a portfolio of classwork as an alternative assessment of proficiency.

their educators experience. The fear of falling short of “proficiency” looms especially large in schools with populations that are short on resources and difficult to teach.

Two Policies in Education

What happens at the intersection of high-stakes testing and academic tracking? The answer to that question is central to this study. While tracking is intended to provide instruction tailored to students according to their academic abilities, high-stakes testing forces those same students to compete with one another. Students who may be more comfortable in the slower pace, relaxed expectations, and easier work of the low-track courses at the end of the year have to take exactly the same test as the students who learned the same material in an honors or AP environment. Taking an optimistic perspective, high-stakes tests may inspire teachers and students in low-track classes to boost their efforts and achieve at a higher level. On the other hand, high-stakes tests may simply drive an existing wedge more deeply between the academic tracks and thus shortchange lower-track students even more.

In North Carolina, both tracking and accountability policies have received substantial attention from educators, administrators, politicians, and the media. Currently the state offers “two courses of study leading to one diploma” (NC Department of Public Instruction, 2012). The graduation requirements for the “future-ready core” differ from the “future-ready occupational” in that the latter has reduced math, science, and social studies requirements. Middle and high schools break down the Core Course of Study further with designations for regular, honors, AP, and International Baccalaureate (IB). AP and IB classes follow curricula set by external organizations, but the qualities that distinguish regular courses from honors courses are, from my own experience teaching at three different North Carolina high schools, vague. One distinction that all honors classes seem to have is an additional boost to the student's grade point average (GPA), but the precise requirements for earning that boost vary from school to school, teacher to teacher, and class to class.

The division of students into separate academic tracks within the Core curriculum is rarely viewed with suspicion or even skepticism, even though research has shown that tracking concentrates racial minorities and the poor in remedial classes that offer reduced opportunities to learn (Oakes, 2005). Rather, offering large numbers of AP and IB courses is a point of pride for many high schools, one that external evaluators see as a positive mark of distinction and high expectations. Honors classes, by extension, are seen as ways to similarly enhance the learning in a school. Rarely to schools brag about the number of their remedial or "regular" track classes. Schools and districts often justify tracking by giving the power of choosing a track to the student or his or her family. Current models of academic tracking in use in North Carolina schools, as in other states, allow students to choose different combinations of tracks for their courses. Although putting students in charge of their own track placement (possibly with some advice from educators and parents with varying degrees of involvement and awareness) may assuage the guilt of teachers and administrators to some degree, research shows that once students enroll in a low track they rarely ever climb out of it (Yonezawa, Wells, & Serna, 2002). In fact, the current "reformed" version of "flexible" tracking quite possibly preserves inequality in our educational system just as well as the earlier, more rigid system (Lucas, 1999). The second chapter of this dissertation describes research on the implications for equity in the context of such tracking policies.

The Applied Research Center, a racial-justice activist organization in Oakland, California, argues against both high-stakes testing and tracking in its report *No Exit? Testing, Tracking, and Students of Color in Public Schools* (Gordon et al., 1999). The document cites research and school documents to describe how these educational practices work together to lock students of color or those in poverty out of educational opportunities. The cycle begins with standardized test scores, on which White students (and often Asian students) tend to do better than racial minorities and the wealthy do better than the poor. Those test results guide decisions by the teachers, parents, and students, both formally and informally, about course selection and academic tracks. Blacks, Hispanics, American Indians, and the poor end up in lower-track classes, where they then receive fewer opportunities to learn. Lower track students learn less and then earn lower test scores which are then used along with grades to place students in subsequent course levels. Along the way, higher-track students receive the encouragement, content, confidence, and skills that prepare them for the next level of education while their counterparts in the lower tracks become relatively less confident, less interested, and more deficient in academic skills.

I recently found myself with the rare opportunity to ask a school superintendent, school board member, and former governor, all from North Carolina, if they believed that academic tracking contributed to inequality of outcomes. The governor and board member each answered with an unequivocal “no” and remarked that tracking was harmless and likely beneficial in the effort to combat educational inequality. The superintendent was not willing to defend traditional notions of academic tracking, which he said tend to get students stuck in a certain trajectory, but believed that grouping students for classes according to achievement was just fine provided there were ample opportunities for mobility among groupings. More details about these conversations appear in Chapter 5.

Policymakers in North Carolina like the ones mentioned above are far more likely to debate high-stakes testing than tracking. Since the 1990's, North Carolina has been a pioneer in the area of state-run end-of-grade and end-of-course tests. Although the required student testing program has experienced fits and starts, the state is currently developing no less than 90 new tests that will assess student learning and teacher effectiveness (Helms, 2012). Conservatives in the state revel in the notion that these tests will hold public school teachers and students accountable. Liberals, however, are of two minds on the topic. On the one hand, they dislike the way high-stakes tests tend to favor the students and school with the most resources. On the other hand, the tests provide measurable evidence of the inequalities within the educational system. The same test that labels a student, subgroup, or school as "failing" also provides evidence that the state is not fulfilling its obligation to provide "a sound basic education" as mandated in the state constitution and multiple North Carolina court cases (*Leandro v. State of North Carolina*, 1996; Packard, 1997).

Today, students in North Carolina take the End-of-Course (EOC) tests in reading comprehension and mathematics in grades 3 through 8 as well as science in grades 5 and 8. Only three high school subjects, Algebra I, Biology, and English I, are currently administered End-of-Course (EOC) tests, but new tests will soon be implemented (NC Department of Public Instruction, 2012). In the cases of both EOG and EOC tests, a student's raw score is converted into a scale score so that results can be compared across test forms and across years. Students and parents are also provided with a percentile score. No distinction is made among the academic tracks when tests are administered, scored, or analyzed. Consistent with the underlying philosophy of accountability policies, all students are held to the same standards when it comes to North Carolina's high-stakes tests. More information about the nature of North Carolina's test will appear in Chapter 3.

The title of this dissertation implies that high-stakes tests are a "distraction" that contributes to social inequality. In the following chapters I present evidence that high-stakes tests are merely the latest educational reform intended to improve outcomes for all students that will do little to bring disadvantaged students to the average level of achievement of their more-privileged peers. Rather than preparing students for the next level of education, high-stakes tests, I contend, distract disadvantaged students from the lessons that develop the skills and knowledge necessary at the next level of education. Students who are White or of high socio-economic status perform better on measures in middle school, high school, and college, and academic tracking plays a role in the mechanism that maintains this inequality of educational outcomes.

This study uses data made available by North Carolina's school accountability policies to explore three questions: (1) do the results of high-stakes tests reflect achievement gaps related to race and socio-economic status, (2) do the gaps grow or shrink over students' educational careers, and (3) how does academic tracking influence the effects of high-stakes tests? The first question is nothing new to the literature on education policy, but it seems necessary to establish that achievement gaps exist before exploring their natures. The second question gets to the issue of schools' capacity to reduce the inequalities associated with student background. Some scholars believe that the significance of an individual's family background fades as the person gets older and has more opportunities to make decisions separate from his or her parents (Müller & Karle, 1993), but other scholars have argued that the effects of background not only persist but increase as a student progresses through schooling, at least in part because family background interacts with school structure to create a snowball effect (Lucas, 2001). The third question takes the debate over high-stakes testing into new territory by considering its impact in conjunction with academic tracking. While plenty of research has established the different ways in which high-stakes testing impacts different schools, very few studies have considered the different effects of academic tracks. Those that do address this question tend to be qualitative in nature. They offer case studies or interviews that show teachers respond to high-stakes accountability policies differently depending on who their students are. Even the same teacher will respond to high-stakes tests differently according to the academic level of his or her classes within the same school day (Watanabe, 2008). This dissertation takes a quantitative approach to that same question and provides statistical evidence supporting the idea that high-stakes tests and all the classroom dynamics they trigger over the course of a semester drive a wedge between upper- and lower-track classes in terms of the learning that occurs in each one. Through the application of multi-level cross-classified models to a unique longitudinal dataset, I show that achievement gaps associated with race and SES exist in middle school, high school, and college, that those gaps increase as students progress through each stage of education, and that North Carolina's high-stakes tests disguise inequalities of educational outcomes associated with academic tracks, at best, and at worst exacerbate those inequalities.

Six chapters follow this one. Chapter Two presents a discussion of the literature relevant to these research questions. I draw from a wide range of sources that analyze the effects of high-stakes accountability policies, the impact of academic tracking on the equity of educational opportunities, and the theoretical foundation of my arguments. Chapter Three lays out the methods used in this study and includes a description of the data and variables used as well as the statistical tests used. Chapter Four presents results of those statistical tests and Chapter Five analyzes those results. Chapter Six describes qualitative studies that support the causal mechanism that I believe explains my results. Finally, Chapter Seven provides a conclusion to the study.

CHAPTER 2: HISTORY, THEORY, AND RESEARCH ON HIGH-STAKES TESTING

This dissertation investigates the effects of high-stakes testing and academic tracking on the gaps in academic achievement associated with race and social class. The previous chapter provided a rough description of the importance of this issue. Sorting students into different academic tracks is a common and firmly-established practice in schools that provides students with different opportunities to learn and frequently separates students by race and class. This form of segregation provides the context in which high-stakes testing operates. This chapter identifies some of the important policies, theory, and research related to the development of high-stakes testing policies and the academic setting defined by tracking.

The Debate over Accountability

By some measures, the United States is a world leader in education. Our literacy rate is high, the number of years an average American has spent in school is high, and every semester many thousands of students from around the world eagerly enroll in American colleges and universities (Clotfelter, 2010). We must be doing some things right. All the same, our education system has plenty of room for improvement. Other countries have shown gains on international assessments of student learning such as the Progamme for International Student Assessment[2] (PISA), and now the United States

[2] PISA tests, developed and implemented by the Organisation for Economic Co-operation and Development (OECD), assess the learning of 15 year-olds in more than 65 different countries every three years in reading, math, and science. The results have been influential on education policy (Hanushek & Woessmann, 2010).

finds itself in the middle of the pack among industrialized nations. However, if PISA test results are disaggregated by race, White students on average earn scores in the top of the international range while Black students score much lower (Darling-Hammond, 2010). In a country that prides itself on both individual achievement and equal opportunity, Americans are worried that our public education system is delivering neither to many of our students, or at least not as much of each as it should.

This study looks at the current widespread strategy for improving achievement and equity: accountability through high-stakes tests. Such accountability policies are premised upon the notion that standards-based standardized tests administered to students can accurately assess the performance of students, teachers, schools, and school districts (Carnoy, 2003). By providing the results of the tests to the public, parents and government officials can supposedly make more informed decisions about schools. Current policies can be traced back to *A Nation at Risk*, the 1983 government report that stoked fears over both education and the economy. In sum, the document argued that American schools were not performing well, students were not learning, and the country's economic competitiveness was waning as a result. A push for accountability followed, including several states instituting accountability programs, and culminating in the 2001 federal law *No Child Left Behind* (NCLB). That act moved federal role in education far beyond what had existed under 1965's Elementary and Secondary Education Act (ESEA), which NCLB reauthorized in a new form. The new law required states, districts, schools, teachers, and even students to demonstrate through test scores that the federal and state governments were getting results for their money (Bush, 2001). This policy was a boon for the testing industry, which was quick to develop the newlyrequired standardized tests that virtually every public school student would need to take to satisfy NCLB regulations (Olson, 2004).

Generations of public school students have taken standardized tests, so initially many Americans were perfectly comfortable with their proliferation, and perhaps even viewed it as a necessary step in the proper evolution of effective schooling. Phrases such as “use only number 2 pencils” and “completely erase any stray marks” and “do not go on to the next section until told to do so” are firmly rooted our memories, have become punchlines for humorists, and sound almost like snippets from a national rite of passage. We adults all survived our Iowa or California Achievement tests, so why wouldn’t today’s kids manage as well?

Moreover, several states had already put standards and accountability policies into effect before NCLB became law. North Carolina, for example, established its ABCs of Public Education program under which the state began sanctioning schools based on student academic performance as early as 1997 (Ladd & Zelli, 2002). Around the same time Texas was using its Texas Assessment of Academic Skills (TAAS) to evaluate and sanction students and schools (Hoffman, Assaf, & Paris, 2001). These and other states’ programs served as models for NCLB and contained requirements that exceeded those of the federal law (Costrell & Peyser, 2004).

Conservatives liked the idea of accountability policies for many reasons, but perhaps chief among them were the opportunities to hold teachers (and their unions) responsible for student outcomes to sharpen the focus of schooling on the academic curricula. Liberals found things to like as well, particularly the generation of data that could be used to monitor and address gaps in achievement correlated with race and social class (Scheurich et al, 2000).

But standardized tests under NCLB and state accountability programs are not the same tests administered to elementary and secondary students from the 1950s to the 1990s. They differ in two important ways. First, the tests are specifically written to match state academic standards, documents with which the state expects teachers to align curricula. Rather than assessing general skills and knowledge, these exams test whether students have learned what the state believes they should learn in a specific course in a specific grade in a specific year. The strict alignment offers several benefits, including a measurement of actual achievement rather than potential achievement and the signaling to students that what they have learned in class is important (Atkinson, 2002). Proponents of such tests welcomed the shift away from the emphases on multiculturalism and student self-esteem that had gained momentum in previous decades in favor of a core academic curriculum (Ravitch, 2001). The second major difference concerns the use of the test results. Under accountability policies, standardized tests take on high stakes. A student’s grade, promotion, or graduation depends on her performance on high stakes tests (Greene, Winters, & Forster, 2003). A teacher’s performance evaluation, job assignment, and possibly even compensation depend on his or her students’ test scores. Likewise, administrators need students to get high scores to protect their budgets and jobs. The strategy is to use top-down control over students, teachers, and schools to change behaviors to improve student achievement (Moe, 2002).

Some educators and education analysts observed a variety of improvements coming out of the high-stakes testing movement. Massachusetts posted gains in test scores that some attributed directly to the high standards and high stakes in its accountability policies, for example (Reville, 2004). Elsewhere, test score results were shown to help overcome the “deficit thinking” among administrators that can be damaging to the opportunities to learn for child of color or from low socio-economic (SES) backgrounds (Skrla & Scheurich, 2004). Wiliam’s (2010) study found that high stakes testing policies do increase student performance as measured on other, lower-stakes tests, but also noted that decision-makers should be cautious not to make too many inferences based solely on test results. One of the most celebrated accountability success stories, that of the “Texas miracle”, was roundly criticized and widely discredited. Valenzuela (2005), for example, showed that the gains in test scores in Texas were made possible by the large numbers of students who dropped out, transferred, or went to jail. While most schools have at least a few students who take such paths, high-stakes tests created incentives for educators to nudge low-scoring and difficult-to-teach students toward them.

Some supporters of accountability policies acknowledge that the methods currently in place are far from perfect. Rothstein, Jacobson, and Wilder (2008) point out that in the effort to keep costs down, government has erred in attaching too many assumptions, conclusions, and determinations to the results of high-stakes tests. Doing accountability right will be awfully expensive, they explain, especially at the beginning. It will require far more comprehensive assessments than we currently use. For now, the systems of accountability currently in use carry with them significant dangers for both students and educators.

The Hazards of Accountability through Testing

McDermott's (2007) historical case studies of accountability policies in four different states indicate that lawmakers instituted high stakes testing with the best of intentions. Frequent, standard-based assessments of students, it was believed, would be a fair and efficient way to inspire improvements in both educational equity and overall achievement. In each of those states, however, the results did not match the policy intentions. Resources did not become more equitable as a result of accountability policies, and rather than demonstrating improvements, low-performing schools and districts shouldered more blame. What is it about accountability that it not only fails to improve things but in some cases makes things worse?

Early in the accountability movement, Darling-Hammond and her colleagues (1995) argued that an emphasis on standardized tests would discourage higher order thinking activities and increase teachers' use of test-specific activities such as recall and recognition. Years before NCLB took effect these researchers observed measurable declines in critical thinking, analysis, and problem solving in classrooms where standardized tests were used. Standardized tests do not capture students' deeper thinking, creativity, performance on a range of tasks, writing, persuasiveness, application of math or other academic skills and concepts, reasoning, or ability to make meaning of data or passages (Medina & Neill, 1988). In order to achieve their purpose of comparing students across schools, districts, and states, standardized tests can only capture a narrow set of skills, which may explain why so few countries, even among those with education levels similar to that of the US, put much emphasis on standardized tests (Darling-Hammond, 2010). When we compare standardized test results to how individuals perform in terms of their lifetime earnings and on-the-job performance evaluations, the correlation is remarkably weak (Levin, 2001), suggesting that the important tasks of schooling are far broader than the qualities that can be captured by standardized tests.

If standardized tests are by their very nature limited in their ability to assess educational achievement, attaching high stakes to them can be downright dangerous, warn some scholars. Linn (2000) cautions that tests that would otherwise be dependable and credible are not once policy applies high stakes to their outcomes. Known to physicists as the Heisenberg Uncertainty Principle and to economists as the Lucas Critique or Goodhart's Law, the rule is familiar to social scientists as Campbell's Law (Campbell, 1979), and states that whenever decision-making is based upon some sort of quantitative social measure, the measure itself will become corrupted and lose validity. Just two years before NCLB was passed, Heubert and Hauser (1999) argued that attaching high stakes to tests causes schools to narrow the curriculum, reduce instructional quality, dilute learning, increase dropouts, reduce graduation rates, encourage or engage in cheating, and favor some students over others. Jones, Jones, and Hargrove (2003) similarly found that high stakes testing policies are associated with reduced curricula, poor teaching practices, reduced motivation, increased retention, degraded reputations of schools, more difficult teacher recruitment, and disproportionate harm to special populations.

Berliner (2011) argues that these consequences reflect completely rational decisions by teachers, administrators, and students. An ambitious educator acting in the interest of himself and his employer would be reasonable to consider engaging in excessive test preparation or even cheating to "game the system." Of course, doing so damages the morale of both teachers and students and causes teachers to engage in educational triage, in which they teach some students more than others – typically the “bubble kids” who are on the verge of making proficiency get all the attention while the students who are likely to pass or likely to fail are not considered worth the teacher’s effort (Nichols & Berliner, 2007).

Empirical research supports these claims. Amrein and Berliner (2002) studied 18 states to find out if the introduction of high-stakes testing had a measurable effect on student learning. The researchers compared average scores before and after policy implementation in terms of ACT, SAT, NAEP, and AP test scores and found that scores declined or stayed flat more often than they increased. More alarming than the failure of high-stakes tests to improve performance on these other measures, Amrein and Berliner found that high-stakes tests prompted an increase in high school dropouts. Similarly, Haney (2000) and McNeil (2005) showed that the “Texas miracle” was made possible by the dropping out, “pushing out”, transferring, incarceration, or deportation of low-achieving students, who happened to be disproportionally Black and Hispanic. In Haney’s study, the high-stakes TAAS test scores showed remarkably weak correlations with one another and with related grades in school, suggesting that the test scores were not reliable. McNeil compared TAAS scores to student scores on TASP (the Texas test for college readiness), SAT, and ACT and found that not only were the correlations weak but that numerous students with good scores on the TAAS were not ready for college when judged by the other assessments. Braun, Chapman, and Vezzu (2010) examined 8th grade NAEP scores in math from 1992 to 2007 across 10 different states to see if the introduction of NCLB improved scores and closed achievement gaps. They detected only very modest gains for Blacks and consistent achievement gaps. Additionally, they found no correlation between the aggressiveness of a state's accountability policy and changes to its test results.

Exacerbating, Not Closing, Achievement Gaps

Arthur Pearl (2002) points out that accountability policies, as currently designed, are based on flawed logic. He states, "testing does not alter life chances any more than measuring temperature reduces fever. In the haste to do something there has been no serious effort to distinguish standards from obstacles." The question is whether testing influences learning. Accountability policies essentially tell teachers and students where their achievement ranks on a specific measure compared to other teachers and students and then to "do better or else." But the ultimatum does not equip schools to overcome the disadvantages that many students must face at the same time they are trying to prepare for a standardized test that will compare their academic progress to that of every other student in the state.

There is little doubt that differences in learning opportunities and obstacles are to a great degree responsible for the "achievement gaps" that appear time and again in education assessments between Whites and Blacks, Whites and Hispanics, and high-SES children and low-SES children. Darling-Hammond (2004) has outlined those concerns, focusing on the disparity of resources across schools that are held to similar standards, and how those disparities are reflected in gaps in achievement. These gaps are already in place when students enter kindergarten and typically persist through high school (Jencks & Phillips, 1998).

Many factors seem to create and maintain the gaps, and a few of the factors are within educators' control. For example, Ferguson (1998) shows that teachers' perceptions, expectations, and behaviors towards white and black students differ, creating educational opportunities that vary with race even within the same classroom. Valencia (1997) found that "deficit thinking", the belief that students of color or low-SES have less potential than other students, on the part of educators is rampant. McNeil (2000) argues that accountability policies are contributing to, rather than reducing, achievement gaps associated with race and class. She supports her argument with the case of Texas, where she finds high-stakes testing accountability policies deskilled the teaching profession, demanded a "test-prep noncurriculum", masked inequalities with dubious test scores, and re-stratified access to knowledge, all by forcing high-poverty schools to compete with low-poverty schools on standardized tests.

More than a decade ago, when the accountability movement was gaining momentum, Ferguson (1998) assessed the potential that various interventions have for closing the achievement gap between Whites and Blacks. Briefly, he found the most promise in three areas: addressing the gap that exists before school, correcting the gross difference in instruction that upper track and lower track students receive, and improving the quality of teachers. Other solutions such as shrinking class size and offering student-by-student interventions, he thought, may yield some improvement but would be far more costly than the other strategies.

Is accountability consistent with Ferguson's three areas of promise? Could high-stakes testing finally be the solution to make educators get serious about raising the achievement of students who are poor and/or black? That was part of the sales pitch for

NCLB given by the Bush administration. In one of the most memorable lines of his presidency, George W. Bush argued that his opponents who were skeptical of the ability of schools and students to improve under the system of NCLB consequences were revealing their “soft bigotry of low expectations.” In other words, the administration was accusing liberals of the very same deficit thinking they claimed to be fighting against. What Bush and other accountability advocates were saying was that students and teachers are not currently living up to their potential because the current system did little to demand that they work hard. Standardized tests, the reasoning goes, would not only expose the districts, schools, teachers, classrooms and students that were performing poorly, but could also provide the basis for a system of rewards and punishments, both symbolic and financial (Kafer, 2004).

Things did not work out that way, at least not very often. As Harris and Herrington (2006) report, once accountability policies came into effect, the achievement gap actually started to grow. The problem, the researchers argue, was not with accountability *per se*, but with the way in which it was being implemented. An accountability system that exposes students to more resources and content might generate better success, but the versions that have been put into effect in the NCLB era have done more to deny students resources and content than to expand them.

One year before NCLB became law, journalist Peter Sacks railed against standardized tests in his book, *Standardized Minds* (Sacks, 2001). He argued that while standardized tests sound egalitarian and democratic – everyone takes the same test and is held to the same standard – they are in fact just the opposite. Sacks believes that our “meritocracy”, in which supposedly anyone can achieve the American dream if she applies herself through schooling, is mostly an illusion. The privileged have certain advantages and use those advantages, according to Sacks, to protect their privileges in the future. Standardized tests allow those in the upper classes feel better about their position because they buy into the idea that standardized tests are completely neutral when it comes to a person's skin color or family income.

Sacks was not alone in his accusations against accountability reforms. Tobin, Roth, and Zimmerman (2001) argued that national standards contributed to "hegemony" reflected in the Black-White achievement gap in their ethnography about a novice science teacher in an urban school trying to make the standard curriculum relevant to students. In her book on the experience of Mexican-American students and their experiences in public schools, Valenzuela (1999) warned that aligning classroom lessons to a standard curriculum that emphasizes the experiences of the majority amounted to "subtractive education" which marginalizes individual students' cultures and thus expands the observed gap. Orfield and Wald (2000) claimed that while standardized tests can provide useful information as formative assessments, using them as the foundation for accountability policies removes their educational benefit and disproportionally burdens Latino and Black students, who are more likely to attend schools with fewer resources and live in communities with less support. Leonardo (2007) argued that NCLB itself was an example of policy that in an attempt to appear race-neutral permits the advantages and disadvantages associate with race to persist unaddressed.

Sacks' book spelled out the concerns over one issue in education policy for a popular audience. Around the same time, sociologist Samuel Lucas described his theory of Effectively Maintained Inequality (EMI). The theory captured a lot of the arguments that Sacks was making, including the false notions of meritocracy and the persistent ability of wealthy families to use their privilege to secure education advantages for their children. A thorough explanation of EMI appears later in this chapter.

What the ideas of Sacks and Lucas have in common concerns the manner in which educational, and thus economic, opportunities are distributed, thus leading to substantial differences in student outcomes. Families play significant roles in how students are introduced to and guided through their schooling (Lareau, 2003). A study by Lee and Burkam (2002) estimates strong correlations among race, SES, family structure, educational expectations, and cognitive ability before a student even begins kindergarten. Lee and Burkam also note that low-SES students begin their schooling in “systematically lower-quality schools than do their more advantaged counterparts.” That concentration of poor and minority children in schools negatively affects their outcomes, as evidenced by the work of Hanushek, Kain, and Rivkin (2009) and Orfield and Lee (2005). In short, the advantages and disadvantages associated with family background matter, and the concentration of students in homogenous schools magnifies the differences in educational outcomes.

Blaming the Victims and Making them into Statistics

The expectation that schools would successfully repair social inequalities if only they had sufficient incentive to do so did not begin with the high-stakes testing movement. As Amy Stuart Wells (2007) points out, Americans have long looked to public education to be the solution to social inequality. While other countries developed social safety nets for people at all stages of life, the United States chose instead to emphasize education, and allow those who perform poorly in schools to figure things out for themselves. The result, Wells points out, is that schools are expected to solve problems that go far beyond the K-12 experience. To make matters worse, schools are expected to fix things during a time of growing income inequality, increased immigration, ongoing segregation, and welfare reform. McDermott (2007) made a similar criticism of accountability by saying that the schools and students who will suffer most under such policies are already suffering, and have little power to overcome their obstacles. The research of Logan, Oakley, and Stowell (2008) shows how unequally educational opportunities are distributed across the landscape of public education in America. Using the city of Boston and its surrounding suburbs for their analysis, the authors found that Black and Hispanic students attended schools that were far more segregated than White students, even when accounting for differences in income.

One lesson that every aspiring teacher learns is the importance of making class content relevant to the students. Good teachers have all kinds of ways of forging those connections, and students generally respond positively (Howard, 2001). High-stakes tests raise serious obstacles for the execution of that basic educational principle. Standardized tests, by definition, must be uniform, and thus they promote a “uniform and objectivist way of knowing and learning to the detriment of cultures, languages, and knowledge” (Valenzuela, 2005). Brenda Townsend (2002) points to three ways in which high-stakes tests harm Black students specifically: the tests ignore racial identity, damage self-concept, and assume a particular achievement orientation. Standardized tests, in other words, go against a lot of what we know would benefit black students most.

Setting aside for a moment the nature of standardized tests themselves, students of color are more likely than white students to face a host of obstacles that hinder learning and academic achievement. In their book *Minority Report*, Gunn and Singh (2004) argue that as a group, students of color need to work harder than white students because individually they are more likely to experience family poverty, neighborhood poverty, violence, neglect, abuse, exposure to drugs or alcohol, poor schools, welfare programs, natal health issues, poor nutrition, single-parent families, foster care, young parents, disillusionment, and low confidence. These students tend to be concentrated in schools that have been abandoned by middle class and white families (Logan, Stowell, & Oakley, 2002), leaving them with frustrated and resource-poor teachers. As results of tests become publicized and parents of prospective students use tests as a proxy for school quality, these schools tend to decay even faster.

Wayne Au (2009) warns that as the curriculum narrows to serve the objectives of high-stakes tests, diversity in the classroom is increasingly ignored by the teachers. In fact, diversity becomes a liability to schools under a high-stakes testing accountability program, because a standardized test will not address the experiences of minority groups beyond the ways in which the majority also experienced or learned about them. To keep up with standards, schools put increased pressure on their minority students, yet at the same time become less responsive to their unique needs. Au gives credit to Grundy, who observed this dynamic a quarter century ago: “student voice and power is increasingly structured out since they have reduced control over determining their own educational objectives” (Grundy, 1987). Johnson (2009) agrees that one of the devastating effects of high-stakes tests is the depersonalization of education, which turns students into statistical objects in order to “maintain the hegemony of the high-stakes testing system.” Her examination of the failed effort to amend the Texas accountability laws to allow for teacher evaluations to supplement test scores led her to conclude that the term “scientific” has become extremely powerful in political discourse, but at the same time its use has strayed more and more from its actual meaning.

To summarize, accountability policies demand that schools correct the consequences of injustices that manifest as achievement gaps and have roots extending far beyond the scope of schools’ power. But the tools meant to solve the problems are in fact making them worse. To educational theorists Bowles and Gintis (1976, 2002), the reproduction of social inequality as a result of education reforms should not be a surprise. In their view, schools are not only mirrors of the division of labor and authority structure in the economy. By design, schools’ main function is to serve capitalist production and reproduce, rather than rectify, the division of labor and social inequality. While *A Nation at Risk* was predicated on the notion that economic conditions are the result of school conditions, Bowles and Gintis argue the opposite is true: economic conditions determine school conditions.

Not many Americans are comfortable with such a depressing, let alone Marxist, view of the institution of public education. The idea of schools as the great equalizer, the doorway to opportunity, and the path of equal access is much more pleasant. But as Michael Apple (1995) points out, Americans ask schools to play two very contradictory roles. On the one hand, public education has a very egalitarian and democratic purpose in reflecting “ideologies of equality and mobility” and satisfying the diverse interests of many different groups. On the other hand, schools must prepare students for a labor market that is very hierarchical and increasingly unequal.

In this section I have outlined the arguments many scholars have made explaining why high-stakes testing, although intended to correct inequalities in education, actually make them worse. These arguments share a firm theoretical foundation, which I describe in the next section.

Theoretical Grounding

In *Tinkering toward Utopia*, David Tyack and Larry Cuban (1995) examine school reforms of the 20th century. They observe that while a multitude of reforms have been proposed and implemented, the core of the American school experience has remained largely unchanged, and thus inequalities have persisted. The authors advance several explanations about how it can be that after so many proposals, debates, and changes our schools still follow the same model of one hundred years ago. One is simply the depth to which traditional schooling has become entrenched in the minds of Americans. Another is the extensive framework of activities that have developed around the familiar school routines that would be disrupted if major reform ever took place. Another explanation is that the politics of education have increasingly lacked a "pluralistic conception of the public good." Reforms in the subsequent decade and half continue to reflect this pattern, with tests aimed at labeling the best schools, best teachers, and "best" students so that those communities can benefit from their success and the others can be punished with lower property values, reduced funding, and shuttered schools. The purpose of such reforms is usually to inspire improvements in achievement, but they also serve to distinguish a "good" education from a "poor" one.

This last explanation touches on the divergent and competing interests at stake in education policy. Frederick M. Hess (1999) has theorized that the frequent but ineffective attempts at education reform observed by Tyack and Cuban are due to policy churn. Professional educators, administrators in particular, can impress voters, parents, government officials, and donors with promises of reforms that work around the edges, but stop short of systemic change. Hess labels such changes “status quo reforms” because the overall structure and operation of schools remains the same (Hess, 2006).

Henig, Hula, Orr, and Pedescleaux (2001) offer another theoretical explanation. Their efforts to understand why even cities with Black leaders have such poor educational outcomes for Black students led them to consider both cleavages in the Black community and the variety of diverse stakeholders in urban education reform. Parents, politicians, educators, and local businesses all have vested interests associated with schools. Too often, the authors argue, these groups operate separately and even competitively. At best, any cooperation among them tends to be in the form of loose connections and fleeting periods of teamwork. The result is a lack of “civic capacity” coming from formal arrangements among interests that build trust and cooperation and might lead to aggressive systemic reforms.

These theories offer convincing explanations for why so many “reforms” turn out to be passing fads, heralded by administrators and politicians but never actually changing the structure or institutions of public education let alone improving outcomes. They have little to say, however, about the few education policy changes that have drastically altered schools and the way students are taught yet still fail to produce more equitable outcomes. The adoption of standards, the growth of high stakes testing, and the ensuing efforts to increase accountability are not the sort of window dressing that Tyack and Cuban, Hess, and Henig address with their theories. Many students, teachers, and parents would argue that k-12 schooling is a different experience than it was a decade ago due to the wildly increased emphasis on standardized tests, strict devotion to state standards, and threat of restructuring or closure in low performing schools.

Whereas Tyack and Cuban, Hess, and Henig argue that education reforms rarely result in actually changing the structure and operation of public education, Samuel Lucas (2001) developed a theory called Effectively Maintained Inequality (EMI) that explains why such significant policy changes occasionally do get adopted but always fail to improve outcomes for the low-achieving students they are intended to help. Briefly put, EMI states that no matter what reform policy makers manage to establish in public education, privileged families will always find ways to secure advantages for their children, thus maintaining social inequality even in the face of reforms. A thorough discussion of EMI should begin with a consideration of some of the social functions of public education.

Schools are not only places for students to learn facts, develop skills, and become socialized. Schools also assign labels to each student, including “dropout”, “honors”, “special ed”, “valedictorian”, and more. The names of schools are signals as well, serving as convenient indicators of the abilities, potential, background, social capital, and status of the people enrolled. It would be nice to think that when a student arrives in kindergarten she has the very same opportunities before her as any other new student. However, rather than overcoming differences in family background, schools tend to maintain them (Jencks & Phillips, 1998). The schools and classes that students from privileged families attend are in many ways superior to the ones that other students attend, and the result is that on graduation day initial differences in social status are the same or perhaps even greater. In the rare instances that students of privilege are integrated in schools with less privileged students, upper class students and families still find ways to come out ahead, and not just through hitting the books and studying hard. This perspective explains, at least in part, why Tyack and Cuban observe so many reforms with so little impact: families with political clout do not want to change the educational system from which they have benefited.

Lucas's Effectively Maintained Inequality (EMI) is a relevant theory for each of the research questions posed in this study. EMI holds that as a particular level of education becomes common, privileged groups will try to maintain their advantages by securing better quality versions of that level of education (Lucas, 2001). As high school completion became nearly universal, the advantaged needed a way to distinguish between the high school educations of their children and those of other groups. Higher tracks, better-credentialed teachers, and better test scores provide those distinctions. A high school transcript that sets an advantaged student apart from her peers by virtue of indications of honors and AP classes will provide that student an advantage in pursuing the next level of education because such labels are seen as an indication of academic merit. Higher track classes also are qualitatively superior to regular classes because they tend to have better teachers, better pedagogy, and higher-status content (Oakes, 2005). Lucas's (1999) own research suggests that when Blacks enroll in higher track classes the improvement in their academic outcomes can be attributed to a combination of the higher aspirations, higher achievement, and more challenging requirements they experience in that course. As if those upper-track advantages were not enough, the "weighted GPA" raises the grade point average of students taking honors classes above those of their peers, usually by as much as a full letter grade (Attewell, 2001).

EMI differs from other theories about the role of student background on achievement in its prediction that the effects of background will continue to grow even as students progress through levels of schooling that are “saturated”, or attended by all social groups. In contrast, the Life Course Perspective maintains that over time a student’s own decisions will become more frequent and more substantial, separating him or her more and more from the effects of family background (Müller & Karle, 1993). Based on that perspective, a student’s background has the greatest impact on educational opportunities and achievement early on, and less as the student gets older. There is a lot of evidence, however, that the effects of family background continue to be felt long into a student’s academic career. An alternative theory, Maximally Maintained Inequality (MMI), explains the strong effects of background on later educational attainment as a result of privileged parents securing more years of education (e.g. college and university) for their children (Raftery & Hout, 1993). MMI states that these class differences are apparent when a level of education is available to only a limited segment of society and that when a level of education is universal, such as high school, class differences are not as pronounced because that level of education is available to everyone. EMI differs from MMI in that it predicts background to play a significant role even within levels of education that are universal. Privileged families, according to EMI, will find ways to make their children’s educations qualitatively better than other children’s, even if there is no quantitative difference in terms of the amount of education.

An important aspect of EMI is the understanding that the advantages provided by family background do not occur accidentally. Privileged parents make conscious decisions to use their resources to gain and preserve educational advantages for their children. Wells and Serna (1996) studied ten schools and identified four specific strategies used by elite families to block efforts of "detracking," or the replacement of tracked classes with heterogeneous classes. Similarly, Welner (2001) described the disproportionate power over school policies that upper middle class families exercised in their effort to derail the court-ordered detracking process in four different schools in four different states.

Achievement tests are another mechanism by which educational experiences at a saturated level of schooling can be shown to be qualitatively different. The dropout rate for Blacks was at its lowest in the 1980s, when schools were the most integrated (Darling-Hammond, 2004) and the achievement gap between Whites and Blacks was smallest at that time as well (Gamoran, 2001). With so much progress being made in terms of equity, EMI suggests that advantaged groups needed another way, besides school segregation, to make their diplomas superior to other groups'. The renewed interest in achievement in the 1980s provided an answer: standardized tests. The advantage that whites and the middle class have on standardized tests such as the SAT has been established repeatedly (Geiser & Santelices, 2007; Geiser & Studley, 2002; Rothstein, 2004). High-stakes standardized tests show similar patterns and have served to widen the perceived gap in achievement between Whites and minorities (Brennan, Kim, Wenz-Gross, & Siperstein, 2001). Even when schools with concentrated minority or low-SES student populations do well on high-stakes exams, the success comes at the price of neglecting the teaching of other important lessons that cannot be assessed on standardized tests (Perna & Thomas, 2009).

EMI suggests that while the standardized tests of accountability policies may be blind to a student's background, the opportunities to learn and develop will likely differ in a manner that correlates with background characteristics. Under high-stakes accountability programs such as NCLB, schools are under pressure to raise both average scores and the lowest scores so that the school can demonstrate growth and proficiency, respectively. Administrators and teachers know the format and content of the standardized tests, and they adjust instruction to maximize scores (Nichols & Berliner, 2007). Because the high-achieving students will most likely put forth good effort and score well, honors classes will spend some time on test preparation but otherwise will operate as they have in the past. Standard or lower-track classes, populated by low-achieving students who might be deficient in the necessary academic motivation or skills, will have instructional strategies adjusted more drastically in an effort to maximize scores. These revised strategies include narrowed curricula and drill-and-kill instruction that utilizes only the lowest-order thinking (Medina & Neill, 1988). Parents who are able to do so provide their children with the prior knowledge, academic skills, instructional support, and learning-focused environments will be more likely to get their children enrolled in the honors classes where drill-and-kill is used minimally in favor of more engaging, challenging, and higher-level activities (James, 2009). As mentioned earlier, privileged parents go to great lengths to preserve tracking programs (Wells & Serna, 1996; Welner, 2001). Outside the classroom, children of privileged families also have opportunities to enrich their children's educational experiences with other activities and lessons that go beyond state standards and standardized tests and extend the lessons students learn in school (Lareau, 1987). As a result, when all students move on to the next level of education, the children of privileged families are better prepared for academic challenges because their educations have included a broader and deeper mix of knowledge and skills.

The research questions posed in this dissertation test several aspects of the EMI theory. If the hypotheses suggested by EMI hold true, the evidence will take the form of (1) standardized tests revealing achievement gaps that increase over the courses of students' educations and (2) the failure of standardized tests to reduce inequality. If the data confirm these expectations, Effectively Maintained Inequality theory presents a bleak outlook on current education reform efforts for those who envision them to be vehicles that foster greater equity and excellence in education.

Research on Testing and Inequality

Lucas's EMI theory presents a discouraging picture of school reform. If true, we should expect that high-stakes achievement tests might actually exacerbate the gaps in education outcomes between higher and lower status groups. This section explores existing literature in the areas of high stakes tests through the lens of EMI.

In the case of high school achievement tests, EMI theory would predict that the tests show a greater division in educational performance between whites and minorities and between rich and poor than could be explained by prior achievement because according to EMI a universal level of education will still exhibit growing class- and race-related differences as advantaged groups find ways to reap the most benefits within the existing educational system. Previous research has shown evidence that this is true. A study by Brennan and colleagues (2001) used data from Massachusetts standardized tests and high school GPA to determine which of the two assessment methods are more equitable. The team found that groups within the population that were already faring poorly did worse on the exams, leading the researchers to conclude that the high-stakes application of exams exacerbates the achievement gap between privileged and non-privileged families and recommended that tests be used for diagnostic purposes rather than student promotion. The study was limited to just one grade of assessments and did not account for school characteristics, however. Differences in opportunities to learn at the schools likely played a role in the observed results.

Of course, it is possible that the wealthy and White students actually do learn more in high school, and the results of the study were simply reflecting that phenomenon. If this is the case, we should see that achievement test scores reliably predict later academic success. If achievement tests are accurately reporting the trajectory that students follow in their academic careers, they may indeed be useful tools for educators and college admissions officers to use. On the other hand, if achievement tests are simply a way for privileged groups to make their high school diplomas stand apart from the crowd without actually reflecting any difference in education quality, we would see that achievement tests act as poor predictors of college success.

Case studies by Perna and Thomas (2009) address this very question. The authors found that state-mandated testing creates an atmosphere in high schools that reduces academic preparation, knowledge, and information and lowers graduation rates, particularly at schools with populations of disproportionally high numbers of racial minorities and poor. The authors contend that such tests are not sufficiently aligned with the academic skills necessary in college and distract teachers and students from focusing on more productive skills and knowledge. This finding matches EMI perfectly. In other words, the less-privileged are under the greatest pressure to do well on high-stakes tests, but when they perform well on those measures and go on to the next level of education, they show a deficit in skills and knowledge because privileged groups have in the meantime used their resources to secure more thorough educations. When held to a common standard by high-stakes tests, privileged families are able to supplement children's education to gain qualitative advantages that will pay off in the next level of schooling.

State-designed achievement tests are not the only exams of dubious value for assessing college readiness. Elliot and Stretna (1988) found that the SAT under-predicted the college success of women and Blacks. In other words, males tend to score higher than women and Whites tend to score higher than Blacks on the SAT, but such gaps do not hold up in college performance. More recently the SAT has been shown to correlate with differences in socioeconomic status as well as race (Rothstein, 2004), and to a much greater degree than high school GPA (Geiser & Santelices, 2007; Geiser & Studley, 2002). Zwick and Himelfarb (2011) pointed out that the socioeconomic status of a school, not just the individual student, also influences SAT scores. If standardized tests are intended to be a method of assessing students across schools, communities, races, and classes, experience with the SAT test suggests they will do poor job of overcoming those factors.

Of course, the SAT is supposed to measure aptitude, not achievement. Because achievement tests are tied to the content of courses, they are expected to be better measures of student progress in particular courses and thus serve as a sort of cross between grades and the SAT (Geiser, 2009). Being standardized allows tests to make consistent measures across large samples, while also tying their content to what is taught in class. However, even the designers of standardized tests make repeated warnings that a test score should never be used to make important decisions without considering other measures. The more that tests are used for accountability purposes the less valid those tests are likely to be (Smith & Fey, 2000). A long list of studies reveals several reasons why.

First, assigning high stakes to tests gives students and teachers a reason to try to game the system and earn a score higher than their level of actual mastery. “Outsmarting the test” could take the form of teachers teaching to the test (Perna & Thomas, 2009; Sacks, 2001), principals manipulating the pool of test takers (Roderick & Engel, 2001), or even outright cheating (Ravitch, 2010). Second, high stakes tests, ostensibly intended to inspire achievement, drive some students to quit school (Heubert & Hauser, 1999; McNeil, 2000), which makes a school appear better by raising its mean scores but simultaneously harms the education level in the community by increasing the number of dropouts. Third, as has been discussed above, high stakes tests exacerbate the achievement gap between whites and minorities (Brennan et al., 2001), which then leads to increased sorting into segregated neighborhoods (Kane, Staiger, & Riegg, 2005). Fourth, whenever educators narrow the curriculum, exclude low-performing students, or engage in cheating, the resulting scores are inflated, thus overestimating the academic achievement of the students affected (Nichols & Berliner, 2007). If inflation takes place equally everywhere, then the results still maintain some degree of accuracy, at least relative to other students. But if certain groups have their scores inflated more often than others, say, groups concentrated in lower-performing schools that are under particular pressure to make adequate yearly progress or earn a certain level of recognition in the state or federal accountability programs, those groups may be missing out on lessons, skills, and knowledge necessary for success at the next stage of education, or in the workplace, or to become independent and active citizens. The scores of these students may be high, but at the next level of schooling, the gaps in their abilities will become obvious.

Even if test scores are accurate, the body of information they cover is often not aligned with the skills needed in college, thus leaving students unprepared (Kirst & Venezia, 2004). That misalignment worsens students' chances for college success as pressure to achieve high scores causes high schools and teachers to focus on the content of state-mandated high-stakes tests at the expense of other academic preparation (Perna & Thomas, 2009). Even with mounting evidence of the limitations and dangers of putting too much stock in standardized tests, many voters and politicians cling to the idea that such tests can be objective and accurate and continue to push for their use in school accountability policies (Johnson, 2009).

This dissertation examines student performance in middle school by North Carolina's End-of-Grade (EOG) scores, in high school by the state's End-of-Course (EOC) scores and high school grade point average (GPA), and in college by GPA. If we assume that academic abilities among students are fairly fixed, we might expect that a student doing well at one stage would also do well at the next, and a student doing poorly at one stage would also do poorly at the next. A deviation from such patterns might be attributed to the differences in experiences related to students' educational development. Lucas's EMI theory suggests that gaps associated with race and SES may actually widen over time, as privileged families provide qualitative advantages for their children at each stage making them better prepared than other children at the next. Advantages could take the form of qualitatively different schools or, as Lucas describes in his book *Tracking Inequality* (1999), advantages could take the form of qualitatively different instruction within the same school via tracking.

Kornhaber's (1997) research in the Charlotte-Mecklenburg Schools (CMS) in the 1990s indicates that track placement, even in a "flexible" system, is determined for most students in middle school. CMS administrators were frank in their description of the Academically Gifted (AG) tracks in middle school as the domain of middle and upper class White children whose parents knew the value of the AG track and how to get their sons and daughters into it. The differences in course content and teaching methods were stark, and by the time the children got to high school the regular track students could not keep up with the AG track students, even though "they had the choice" to enroll in the very same honors class.

Scholars Harris and Anderson (2012) explain how the inequalities of tracking might be exacerbated by high-stakes accountability. The authors start with the premise that discourse has proven to be an effective method for teaching mathematics. They argue that in high-track math classes, teachers tend to be quite comfortable engaging students in discourse, thus providing those students with excellent, state-of-the-art instruction. Teachers in low-track classes, however, tend to have less math and pedagogical knowledge, avoid rigor, and limit instructional methods to those reflected on high-stakes assessments. Thus, lower-track classes are denied the opportunities to learn found through mathematics discourse. An example of quantitative analysis that demonstrates the relationships among family background, academic track, and student outcomes can be found in Klugman's (2012) national study of tracking's effects on college destinations. In it, he finds that not only do high school resources mediate the effects of family background on students' pursuit of higher education, but the student's academic track within the high school does as well. If such differentiation in learning opportunities is indeed occurring under current accountability policies, the results of this study should reveal that high-stakes testing, in conjunction with tracking, exacerbates achievement gaps. The ultimate result of this pattern is the reproduction of social inequality.

This dissertation draws from the literature described above by examining (1) the achievement gaps associated with high-stakes tests, (2) the expansion of those gaps over the course of students' educations, (3) and the manner in which academic tracking and high-stakes tests work together to reproduce inequality. In the remaining chapters of this dissertation I test five distinct hypotheses:

H1: White and high-SES students perform better on high-stakes tests than racial minority and low-SES students.

H2: Controlling for prior achievement on high-stakes tests, white and high-SES students perform better on high-stakes tests than racial minority and low-SES students.

H3: Controlling for high school achievement, white and high-SES students perform better in college than racial minority and low-SES students.

H4: Controlling for high school achievement, upper-track students perform better in college than lower-track students.

H5: Upper-track students will perform better at the college level than lower-track students to a greater degree than their high-stakes tests scores would predict.

The next section describes the data and methods used to conduct the analysis.

CHAPTER 3: METHODS

The previous chapters reviewed the existing literature on high-stakes testing, academic tracking, and the effects they have on equity in academic outcomes. Research shows that high-stakes testing causes educators to change their teaching to have a greater emphasis on test preparation, especially for students who are likely to earn low scores. The increased emphasis on test preparation takes the form of rote memorization and other lower-order cognitive skills at the expense of higher order skills such as analysis, synthesis, and evaluation. Lower track classes, when compared to honors classes, are one of the settings where teachers often emphasize test preparation. This dissertation examines whether differences in test performance across racial and social groups are evident in North Carolina's high-stakes tests, whether those gaps increase as students progress through school, and whether a student's academic track affects the correlation between high-stakes test performance and college performance.

This chapter describes the research methods used in this dissertation. The chapter is organized with four parts. The first identifies the sample, the second describes the data, the third defines the variables, and the fourth lays out the analytic steps.

The Sample

The sample used for this project consists only of North Carolina public school students who graduated high school in 2004. In terms of student achievement, student-teacher ratio, and expenditure North Carolina resembles the United States average. Table 1 presents several such comparisons.

Table 1**.** Education measures comparing North Carolina to the United States.

Measure	**North Carolina**	**United States**
Percent at or above basic proficiency in math, 8th Grade, 2005	72	68
Percent at or above basic proficiency in reading, 8th Grade, 2005	69	71
Average SAT score, 2006	1,008	1,021
Percentage of population over 25 with a high school diploma, 2005	84	85.2
Percent of students in private schools, 2003-2004	7	9.6
High school dropout rate	5.7	3.6
Pupil-teacher ratio, 2004-2005	14.8	15.6
Percent of public school students with disabilities, 2005	14.3	14.1
Education spending as a percentage of personal income, 2004	7.2	7.2
Education spending as a percentage of general spending, 2004	35.6	34.4

Table made by author based on Hovey and Hovey (2007).

North Carolina also resembles the United States in terms of racial diversity and SES. In 2005, North Carolina was 68% White while the United States was 67% White.[3] In the same year, 13% of North Carolina residents lived in poverty, matching the rate for

[3] While the majority population in North Carolina resembles the United States, the minority population is disproportionally black.

the United States overall (Hovey & Hovey, 2007). These similarities make North Carolina's population useful in drawing conclusions that may be applicable to many other states.

Perhaps the most important reason to select North Carolina for this study is its role as a pioneer in the accountability reform movement. Along with Texas, North Carolina was one of the earliest states to implement high-stakes testing and receive national attention for increasing student achievement on both state tests and NAEP (Grissmer & Flanagan, 1998). North Carolina's accountability program is called The ABCs of Public Education, a name which stands for "Accountability and high standards, the Basics, and local Control." Between 1995 and 1998 the state implemented the ABCs program for every public school from kindergarten to 12th grade. The program continues to evolve, but the use of standardized test results to evaluate, reward, and punish students and schools has been a part of the ABCs from the beginning.

The state of North Carolina is also useful for a study like this one due to the presence of an affordable, diverse, and geographically dispersed public system of higher education. North Carolina supports both two- and four-year schools; the sample in this dissertation includes only the latter. The sixteen four-year campuses of the University of North Carolina (UNC) system are managed by a single General Administration, making it easy to collect and maintain data on a significant portion of the state's high school graduates in their pursuit of higher education. More details about those campuses are discussed later.

The sample includes the North Carolina's entire 2004 high school class (more than 76,000 students) and the 21,359 students in the subset who went on to the UNC system. Data from high school years include students' test scores, middle and high schools, SAT scores, grade point averages, and family background information. Data from college include UNC campus, grades, majors, and graduation. In terms of race, 67% of the students appearing in the middle school, high school, and college data are white, 25% are black, and 8% are other racial categories; 55% are female and 45% are male.

The sample of students is spread across 16 UNC campuses. Five of the campuses are historically Black colleges and universities and one, Pembroke, has a large Native American population. NC State and Chapel Hill are considered flagship institutions, the former emphasizing science and engineering (as does NCA&T, one of the historically Black institutions) and the latter offering a traditional liberal arts education.

The sample includes data on all EOC and EOG test-takers in the 2004 cohort, not just the ones going on to the UNC system. The nature of the EOC data enables me to investigate selection bias/selectivity in UNC system attendees. For example, the race and gender of the overall sample are reported and compared to the UNC-bound students in Table 2. In terms of race and sex, the two groups are similar.

Table 2. Comparison on demographic characteristics of the students in the sample who enrolled in the UNC system and the sample overall.

	RACE						**SEX**	
	White	Black	Native	Asian	Hispanic	Other	Female	Male
NC public school class of 2004	62%	31%	1%	2%	3%	1%	50.5%	49.5%
NC public school class of 2004 entering UNC	67%	25%	1%	4%	2%	2%	55%	45%

Data

All models in this study use the same longitudinal survey dataset, the North Carolina Roots of STEM dataset compiled by Drs. Elizabeth Stearns, Roslyn Mickelson, Melissa Dancy, and Stephanie Moller in the sociology department at UNC Charlotte. Their Roots of STEM project, supported by a grant from the National Science Foundation (NSF REC-0635004), seeks to understand what factors increase the likelihood that women and minorities will declare and complete majors in the science, technology, engineering, or math (STEM) disciplines. This study is part of that larger project. The dataset follows one complete statewide North Carolina public school graduating class from middle school through high school and then into the state's UNC system. More than 20,000 students graduated from NC public schools in 2004 and matriculated into the UNC system; all of them appear in this data.

The North Carolina Education Research Data Center (NCERDC) at Duke University received the data from North Carolina's Department of Public Instruction (DPI), the College Board, and the UNC General Administration and prepared it for our use by developing crosswalk files with anonymous student identifiers. The data came in parts and were compiled by the ROOTS research team. Individual student–level data came in EOG and EOC datasets based on year, covering 1998 through 2004. Student-level data at the college level includes datasets on the cohort at the time of enrollment, declared majors, financial aid, and graduation. Individual students are identified only by encrypted identification numbers. Other datasets include teacher-level, school-level, classroom-level, and school district-level datasets with variables such as teacher licenses, test scores, demographics, and more. The various datasets were merged by student identifiers so that the researchers have a single dataset with middle school, high school, college, and individual level indicators for each student.

Variables

The analysis in this study focuses on student-level variables, but the models control for the random effects of high schools and college campuses to account for variance that may be associated with attendance at a particular high school or UNC campus (Laird & Ware, 1982). Table 5 provides an overview of the student-level variables. Note that while differences in achievement associated with sex are not a focus of this dissertation, I include sex as a control variable because an achievement gap between males and females, favoring the latter, has been well documented (Orr, 2011), and the difference between males and females in terms of entering college (again favoring females) is likewise well-established (Corbett, Hill, & St. Rose, 2008).

As Table 3 reveals most of the variables used in this study came through NCERDC from North Carolina’s Department of Public Instruction (DPI). Variables for sex and limited English ability needed no recoding. Race variables resulted from breaking apart the categorical variable provided in the dataset and creating a dummy variable for each race. I wanted to capture two aspects of socio-economic status, both parent education and income. The DPI data included a variable for parent education level. I recoded it into dummy variable that reflected whether or not neither of the student’s parents completed high school. The link between education and SES is fairly well established in the literature (Lareau, 2003). The DPI data also included a variable for participation in the federal Free and Reduced Lunch (FRL) program while in middle school.[4] I used these two variables to create a low-SES variable that is coded 1 for any student with received free or reduced lunch in middle school or had parents without high school diplomas.

Table 3. Identification of core terms, student-level variables, and their operationalization in the Roots dataset.

Core Terms	**Variable**	**Operationalization**	**Data Source**
High School Achievement	High School GPA	High School GPA	UNC
	SAT Scores	Total math and reading SAT scores	UNC
	EOC Scores	Mean score of student's required EOC tests (algebra I, geometry, algebra II, ELP, US history, English I, biology)	DPI
Opportunities to Learn	Track Placement	Proportion of EOC courses taken as honors, AP, or IB	DPI
Middle School Achievement	EOG Scores	Mean scale scores for math and reading EOGs	DPI
Student Background	Race	Dummy variables for each	DPI
	Sex	Dummy variable in which male=1	DPI
	Low SES	Dummy variable in which 1 indicates either free/reduced lunch while in middle school or parents did not graduate high school	DPI
	Limited English Ability	Dummy variable in which 1= limited English skills	DPI
College Experiences	College GPA	Quality points divided by attempted credits	UNC
	Declared Major(s)[5]	Dummy variables for each category of declared majors: humanities, social sciences, business, health, professional, education, and STEM.	UNC

[4] Our data do not include information about free or reduced lunch in high school.

[5] To save space, the coefficients for declared majors are not shown the results for all tables, but the variables are included in all models for which college GPA is the dependent variable.

While FRL is far from a perfect indicator of SES, it remains widely used in education research (Harwell & LeBeau, 2010). By combining FRL with a measure of parent education, I attempt to remove some of the bias that may be present in the FRL measure alone. Table 4 illustrates the creation of the Low SES variable.

Table 4: Creating the low SES variable from parent education and FRL measures in the Roots dataset.

Variable	**Observations**	**Mean**	**Std. Dev.**	**Min**	**Max**
Parents did not graduate high school	70603	0.070	0.255	0	1
Free or Reduced Lunch	66471	0.341	0.474	0	1
Low SES	70714	0.323	0.467	0	1

I used DPI data to create measures of mean EOC scores and mean EOG scores. Although North Carolina's requirements for EOC tests are currently in flux, at the time of this cohort's high school education students in the Core curriculum were required to take high-stakes tests in Algebra I, Geometry, Algebra II, ELP[6], U.S. History, English I, and Biology. Middle school students took reading comprehension and mathematics EOG tests in both 7th and 8th grades. All of these tests follow a multiple choice format. DPI converts students raw score into a scale score that can be compared to the student's scale score on the same exam in previous (and later) years. Scale scores are reported to parents, along with a percentile rank score and an achievement level, which fits into one of four pre-determined performance benchmarks.

[6] ELP stands for Economics, Law, and Politics. This sophomore-year course later was renamed Civics and Economics.

Each item on the EOG and EOC tests is directly tied to the goals found in the state standards. For example, Goal 5 of North Carolina's 8th grade language arts curriculum is for students to "use interpretive and evaluative processes to analyze texts and their characteristics" (DPI, 2012). On the EOG, students needed to read a passage from "The Final Memo", a short story by Paul Stewart, and answer five multiple choice questions about characters, their personalities, and word usage in the reading.

Table 5. Calculating the mean EOG and mean EOC scores in the Roots dataset. Each mean test score reflects the student's average scale score on required exams.

Variable	**Observations**	**Mean**	**S.D.**	**Min**	**Max**
Reading EOG	70605	161.0	7.60	131	183
Math EOG	70654	172.2	10.1	135	203
Mean EOG	**70714**	**166.6**	**8.3**	**133.5**	**191**
Algebra 1	62114	62.6	8.7	31	94
Geometry	47583	61.2	9.7	33	101
Algebra 2	42081	65.9	9.7	33	101
ELP	61665	56.0	8.3	21	86
U.S. History	49167	58.2	8.0	31	87
English 1	67558	56.1	8.1	22	83
Biology	57501	58.5	7.3	31	86
Mean EOC	**70714**	**57.9**	**8.3**	**21**	**87**

Table 5 shows the EOG and EOC scale scores used to create the mean EOG and mean EOC variables for each student. To create the mean EOG variable, I averaged students' reading and math EOG scores. To create the mean EOC variable, I averaged students’ required EOC test scores, including Algebra 1, Geometry, Algebra 2, English 1, ELP, U.S. History, and Biology. Table 5 shows that the EOG math and reading scores are scaled to similar means, ranges, and standard deviations. Likewise, the EOC tests are scaled to similar means, ranges, and standard deviations. An alternative approach would be to use scale scores from each EOC test separately. I chose to average the scores instead because other measures, like grade point averages, successfully predict college success while averaging many different areas of study into one variable and include data from multiple points in time. Additionally, using a mean EOC score permits the inclusion of students who may be missing scores on some of the tests. Yet another approach would be to do a factor analysis with the scores, but interpretation of factor results is more difficult than standard regression coefficients.

The EOC data also offers ways to identify students’ academic tracks. All EOC courses have an indicator of course level. From these variables I created a proportion honors classes variable by taking the number of EOC-tested honors-, AP-, or IB-level EOC courses a student had taken among and dividing it by the total number of EOC courses they had taken. Although this method only incorporates a small portion of the high school career, it covers multiple disciplines and the most important courses, if only because accountability policies give them such importance. I believe it is likely that students, parents, and educators give the most attention and thought to track placement for the courses with high-stakes tests. The role of high-stakes tests also makes this variable a conservative estimate of the effects of tracking generally. Because they are state-tested, these courses are especially tied to the core curriculum, so any differences we see as a result of tracking are probably tied to differences in how they are taught and not due to differences in content.

Another possible way to measure academic tracking is by the pace at which students go through the prescribed sequence of math courses. Algebra1, geometry, and algebra 2 were all courses with high-stakes tests attached to them. In math, some students move through the sequence early while others complete it later. Using the pace of math classes presents two problems, however. First, it is highly correlated with the proportion honors variable, because students who take the math sequence early also tend to be honors students. Second, it did not have nearly as many observations as the proportion honors classes variable did. Third, the math sequence frequently bridges middle and high school, with many honors students taking Algebra1 and even geometry in middle school while others wait until high school, making school effects hard to capture. Finally, efficient scheduling often results in mixing different grades and ability levels in the same classrooms, which is a form of detracking, but one that is not likely to follow consistent patterns from school to school and class to class. After using math pacing in several early models, I eventually dropped it.

The discussion of the tracking variable is a good place to address the issue of selection bias. One of the potential flaws in the design of this study is the fact that the only college data included in the sample is for students attending the UNC system. Students enrolling in private institutions or out-of-state public schools or community colleges do not appear in the data. The table below shows the mean EOC scores that students at three different levels of tracking earned, on average, among students in the UNC-bound data compared to those who did not enter the UNC system. The gaps are fairly consistent. At the all-honors track, mixed-level track, and no-honors track, students bound for the UNC system scored about five points higher than students who did not enter the UNC system. At each level of tracking, UNC-bound students out-perform other students on mean EOC scores by between four and six points. While the table indicates that the two groups differ from each other in terms of standardized test performance, that difference appears consistent across three different categories of academic tracking. Therefore, although selection bias may cause the EOC scores to be higher in my analytic sample, that bias is similar across tracks.

Table 6. Mean EOC scores, on average, for students entering the UNC system compared to students not entering the UNC system and compared across different academic tracks.

	All EOC courses at honors level	**Some honors level EOC courses**	**No honors level EOC courses**
UNC-bound	68.0	64.9	58.9
Not UNC-bound	63.9	60.9	53.2

High school GPA data came to our research team from NCERDC as part of the UNC enrollment data. The same was the case for SAT scores and campus of enrollment. Declared majors appeared in the UNC "major" dataset and identified all declared majors for each student. I re-coded the variables by discipline according to the Classification of Instructional Programs (CIP) codes[7] in the data to create seven different variables, each representing a different category of disciplines. A student who declared a major in the humanities received a 1 for the humanities major variable, while a student who did not

[7] CIP codes have been used by the U.S. Department of Education since the 1980s to track developments in different fields of study. The first few digits indicate broad categories of disciplines while the following digits zero in on more specific fields.

declare humanities received a 0. This method allowed students to have majors in more than one subject area and allowed my models to account for differences in difficulty and competitiveness. I developed a variable for college GPA based on the data found in the UNC "grades" dataset and followed the instructions provided by NCERDC. The grades in the dataset reflected each individual course, so I divided each student's quality points by course credits to arrive at a single value to reflect average points per attempted credit.

Some of the questions investigated in this study use only data from middle and high school. Other questions require the use of college data as well. Table 7 provides descriptive summaries for the variables in the models using only middle and high school records. Table 8 provides descriptive summaries that include college records as well. The dramatic change in sample size is explained by the fact that not all high school students go on to college and among those that do, not all enter the UNC system.

As these two tables indicate, there is a substantial divide between number of observations used in the middle and high school level models and the models that include college experiences. I have already shown that although UNC-bound students performed better than non-UNC-bound students on EOC tests, the gap remained consistent across academic tracks, suggesting that the loss of students was distributed fairly evenly across tracks in terms of test performance. The two analytic samples remain similar in several other important respects. First, the ratio of males to females is similar in the two samples. Second, the distributions of racial groups are similar in the two samples. Where the two groups differ most is in SES, which is 32% in the larger sample and 16% in the UNC-bound sample, and honors courses, which averages 22% in larger sample and 46% in the UNC-bound sample.

Table 7. Descriptive summaries for middle and high school analyses.

Variable	**Observations**	**Mean**	**Std. Dev.**	**Min**	**Max**
Dependent Variables					
Mean EOG score	70714	166.609	8.316	133.5	191
Mean EOC score	70714	57.948	8.310	21	87.0
Independent Variables					
Male	70714	0.489	0.500	0	1
White	70714	0.666	0.472	0	1
Black	70714	0.270	0.444	0	1
Hispanic	70714	0.020	0.141	0	1
Asian	70714	0.018	0.132	0	1
American Indian	70714	0.016	0.124	0	1
Other race	70714	0.000	0.018	0	1
Limited English	70714	0.011	0.104	0	1
Low SES	70714	0.323	0.467	0	1
Proportion honors classes	70714	0.220	0.310	0	1

Neither the gap in SES nor the gap in tracking is unexpected. Certainly a greater proportion of honors-track students is going to go into college than from the population overall. Likewise, low-SES students will not enter college at a rate as high as other students. For these reasons, the latter models in this study, the ones that use college-level data, are limited in their generalizability due to selection bias. At the same time, there is reason to believe that the bias is toward more conservative estimates of inequality rather than more exaggerated estimates. My analyses test whether track and SES are factors that result in different educational outcomes for students. If lower-track students and lower-SES are disproportionally excluded from some of my models because they did not have the opportunity to attend college, the actual harm to educational equity is probably greater than that reflected in my results.

Table 8. Descriptive summaries for middle school through college analyses.

Variable	**Observations**	**Mean**	**Std. Dev.**	**Min**	**Max**
Dependent Variable					
College GPA	18647	2.629	0.881	0	4.52
Independent Variables					
Mean EOC score	18647	64.189	6.156	41.7	88
Male	18647	0.439	0.496	0	1
White	18647	0.686	0.464	0	1
Black	18647	0.251	0.433	0	1
Hispanic	18647	0.013	0.112	0	1
Asian	18647	0.030	0.171	0	1
American Indian	18647	0.011	0.104	0	1
Other	18647	0.003	0.057	0	1
Limited English	18647	0.008	0.088	0	1
Low SES	18647	0.157	0.363	0	1
Proportion honors classes	18647	0.456	0.334	0	1
High school GPA	18647	3.61	.660	0	5.53
SAT Total	18647	1072	175	510	1600
Science major	18647	0.179	0.383	0	1
Professional major	18647	0.031	0.173	0	1
Health major	18647	0.056	0.230	0	1
Business major	18647	0.133	0.340	0	1
Humanities major	18647	0.128	0.334	0	1
Education major	18647	0.122	0.327	0	1
Social science major	18647	0.183	0.386	0	1

Analytic Steps

This study tracks tens of thousands of individual North Carolina students through middle school, high school, and college. I develop several models through which I test the effects of student background and prior achievement on academic performance. I control for the effects of high schools through the use of multi-level models. These models allow the effects of individual characteristics to be separated from the effects of schools and college campuses. A standard multiple regression equation includes an outcome, an intercept, a coefficient for each independent variable, and an error term. Multi-level regression models add another term that represents another set of intercept and coefficients for each additional level of clustered observations. The models in this study that include student-, high school-, and college-level data are not only multi-level, but also cross-classified because the high schools are not nested within colleges (or vice-versa). Students from a single high school might enroll in different colleges and students at a particular college may come from different high schools.

All models predict student performance at multiple levels according to the following equations:

Level 1 (Students)

$$Achievement_{ij} = \pi_{oj} + \sum_{p=1}^{p} \pi_{pj}\, a_{pij} + e_{ij}$$

Level 2 (Schools)

$$\pi_{0jk} = \Theta_p + \sum_{q=1}^{Q_p} (\beta_{pq} + b_{pqj})X_q + b_{p0j}$$

The level-1 model estimates the student achievement as a function of student characteristics where

π0j is the achievement for student i in school j;
πpj are the effects of student characteristics on STEM enrollment in school j; and
eij is a random error associated with student i in school j.

At level 2, the school level, the average student achievement is modeled as a function of school characteristics (both secondary school and college) and the other level-1 coefficients are all fixed at their respective grand means, where

Θp is the average student achievement across all schools;
βpq are the fixed effects of school characteristics on student achievement;
bpqj are the random effects of school characteristics Xq;
bp0j is the residual random effects of schools.

By tracking this cohort of students into the UNC system, we can develop an understanding of how inequality at the college transition compares to the high school transition. Because college is less saturated than high school, EMI[8] would suggest that inequality would be greater at the transition into college than the transition into high school.

I began with a two-level analysis of student performance on high-stakes tests in high school and used the software package Stata and its xtmixed procedure run multiple regression models that account for high school effects (Rabe-Hesketh & Skrondal, 2012). These models address the first two hypotheses, which link student background and prior achievement to performance on high stakes tests. To account for collinearity problems that may occur in multilevel regression analysis, I centered the variables by the "grand

[8] The theory of Maximally Maintained Inequality (MMI), put forward by Raftery and Hout (1993), resembles EMI in that it predicts that the effects of social background will increase as a student progresses through school. The Life Course Perspective (LCP), on the other hand, states that the effects of social background should fade as the student gets older (Müller & Karle, 1993). See chapter 2 for a more thorough comparison.

means" by subtracting the mean value of the observations in the analytic sample from each observation. Some students changed schools during high school. I chose to limit my analysis to whatever high school students were in during 10th grade because a student is likely to take more EOCs in the 10th than in any other year.

I included interaction terms in several models to examine the way in which prior achievement may have different effects for Blacks and low-SES students than for Whites and high-SES students. I use these variables to detect significant moderating effects of race and SES with prior achievement on later achievement. EMI theory says that background characteristics, such as race and SES, can become more important as time goes on. An interaction term using race and prior achievement or SES and prior achievement can show how middle school test scores can mean different things for students depending on their race and SES. An interaction term that has a negative and significant coefficient supports the hypothesis that when a student is Black or of low SES, any increases in his or her prior achievement will result in a lower outcome for that group than for middle class or White students making the same improvement.

The other three hypotheses follow students through high school and college, requiring that an additional level of clustered observations be accounted for. Colleges and high schools are not nested within one another, of course, so for these analyses I used the software package SAS and its proc mixed procedure to perform cross-classified multilevel analyses that allow for the effects of both high schools and campuses within the UNC system (Singer, 1998). Independent variables were centered according to grand means in the analytic samples. Again, a decision had to be made about students who transfer campuses during their college career. I opted for limiting analysis to the campus where students enrolled their first year.

The next chapter presents the results of my statistical tests and includes a brief description of each. The subsequent chapter provides analysis of those results.

CHAPTER 4: RESULTS

This dissertation examines inequalities associated with high-stakes testing accountability policies, particularly in conjunction with the practice of academic tracking. Previous research has shown that although accountability policies are intended to raise the achievement of low-performing students, teachers, and schools, evaluations of such policies have been mixed at best. This dissertation examines the gaps in achievement on high-stakes tests between different racial groups and social classes, the shifts in those gaps as students move through schooling, and the relationship between those gaps and academic tracks. To conduct these investigations, I employ cross-classified multi-level modeling and a dataset that follows one complete cohort of students through North Carolina's middle schools, high schools, and four-year public university system. In this chapter I present the results of the analyses conducted to test the hypotheses presented earlier.

High School Achievement

The first hypothesis in this study states that minority and low-SES students will score lower than White and middle and upper class students. The second hypothesis extends that statement to say that even when accounting for previous achievement, the gaps in achievement will continue to grow. Table 9 tests those hypotheses in terms of high-stakes tests scores through the results of multi-level regression models predicting mean EOG and EOC scale scores by student background. Students are clustered by high school to account for school effects. Models 1 and 2 show that low-SES students score, on average, about 3.5 scale-score points lower than other students on both the seventh grade EOG tests and high school EOC tests. Whites (the reference category) tend to score higher than Black, Hispanic, and American Indian students, but lower than Asian students.

Table 9. Multi-level models predicting EOG and EOC scores.

	Model 1: Predicting EOG scores		**Model 2: Predicting EOC scores**		**Model 3: Predicting EOC scores with prior achievement**	
	Estimate	**S.E.**	**Estimate**	**S.E.**	**Estimate**	**S.E.**
Low SES	-3.543*	0.066	-3.439*	0.065	-0.666*	0.039
Black	-5.165*	0.072	-4.462*	0.070	-0.411*	0.043
Hispanic	-0.616*	0.210	-0.087	0.204	0.397*	0.121
Asian	1.486*	0.213	2.901*	0.207	1.744*	0.122
American Indian	-3.395*	0.240	-3.279*	0.234	-0.618*	0.138
Other race	2.002	1.452	3.052*	1.413	1.470	0.834
Male	-0.522*	0.054	-0.332*	0.052	0.087*	0.031
Limited English	-6.381*	0.290	-4.644*	0.282	0.369*	0.167
Mean EOG score					0.787*	0.002
Constant	170.265	0.184	62.264	0.200	59.252	0.085
Log-likelihood	-239283.0		-237368.1		-200049.0	
N	70714		70714		70714	

**Significant at .05*

The third column of Table 9 adds mean EOG scale scores to the model predicting mean EOC scale scores as an independent variable. The inclusion of prior achievement on a similarly-constructed measurement reveals that low-SES, Black, and American

Indian students perform worse on high school achievement tests than their middle school scores would suggest, relative to Whites and students not of low-SES backgrounds. Hispanic, Asian, and male students do better than their previous test scores would suggest, relative to their counterparts.

Table 10. Predicting mean EOC score with interactions and tracks.

	Model 1: Effects on mean EOC score		**Model 2: Effects on mean EOC score with tracking**	
	Estimate	**S.E.**	**Estimate**	**S.E.**
Proportion honors classes			3.430*	0.063
Mean EOG	0.785*	0.002	0.719*	0.002
Low SES	-0.829*	0.040	-0.600*	0.039
Mean EOG * Low SES	-0.065*	0.005	-0.031*	0.005
Black	-0.715*	0.045	-0.697*	0.044
Mean EOG * Black	-0.083*	0.005	-0.062*	0.005
Hispanic	0.509*	0.120	0.471*	0.118
Asian	1.772*	0.122	1.500*	0.119
American Indian	-0.533*	0.137	-0.625*	0.134
Other	1.408	0.830	1.399	0.813
Male	0.073*	0.031	0.177*	0.030
Limited English	0.380*	0.167	0.322*	0.164
Constant	59.002*	0.082	58.880*	0.079
Log-likelihood	-199674		-198209.2	
N	70714		70714	

** significant at .05*

Another test of the relationship between social position and educational outcomes can be done by introducing interaction terms to the model above. A multiplicative interaction term added to the model measures the degree to which the relationship between two variables might be conditional upon some third variable. In this case I want to learn whether the relationship between EOG scores and EOC scores is conditional upon social class or race. Specifically, I examine race in terms of students being Black, because it is the Black-White achievement gap that appears to grow in magnitude in the previous table. Table 10 includes two interaction terms created by multiplying the mean EOG scores by two dummy variables: Black and low SES. The second model in the table adds a variable accounting for the students' academic tracks.

Based on the results of the first model in Table 10, the equation for SES looks like this:

$$\text{Mean EOC} =$$

$$59 + (.785)(\text{EOG}) + (-.829)(\text{SES}) + (-.065)(\text{EOG})(\text{SES})$$

when other variables are held to their means. The equation for race (Black compared to White) looks like this:

$$\text{Mean EOC} =$$

$$59 + (.785)(\text{EOG}) + (-.715)(\text{Black}) + (-.083)(\text{EOG})(\text{Black})$$

Results reveal negative, statistically significant coefficients, showing that not only do gaps exist along racial and SES lines, but also that a higher EOG score during middle school by a student who is Black or low-SES will result in a smaller high school improvement than other students would show. Because all explanatory variables are centered on the grand means, the coefficients for Low SES and Black are the coefficients for those variables when mean EOG scores are average. The maximums for both low SES and Black, after centering, round to .7. In the first model, being of low SES will reduce a student's EOC score by .829(.7) =.580 points when they earn average EOG scores. A Black student with average EOC scores will lose .715(.7) =.5 points. Although statistically significant, these differences do not sound like terribly substantive gaps.

Considering the explanatory variables in conjunction with the interaction term tells more of the story. Other factors held to their means, low SES student has an equation predicting EOC score that looks like this:

$$\text{Mean EOC} = 59 + (.785)(\text{EOG}) + (-.829)(.7) + (-.065)(.7)(\text{EOG})$$

$$= 58.4 + .739(\text{EOG})$$

when EOG refers to the difference between the student's mean EOG score and the average mean EOG score. By contrast, a student who is not low SES, as defined in this study, would have an equation that looks like this:

$$\text{Mean EOC} = 59 + (.785)(\text{EOG}) + (-.829)(-.3) + -.065(-.3)(\text{EOG})$$

$$= 59.3 + .805(\text{EOG})$$

when EOG refers to the difference between the student's mean EOG score and the average mean EOG score. Notice that the intercepts are slightly different; holding other factors at their means, low SES students are predicted to have a mean EOC score of 58.4 while other students are nearly a full point higher. Notice also that low SES students have a flatter slope in their equation. If a low SES and another student who is not low SES both score at the mean on their EOG tests the low SES student will score nearly a point lower on her mean EOC score and for each point of improvement on mean EOG scores the low SES student will trail her counterpart by another .066 points on mean EOC scores.

Compared to Whites, Blacks show a disadvantage as well. The minimum for the centered Black variable is -.29 and the maximum is .71. A Black student's predicted EOC score looks like this:

$$\text{Mean EOC} = 59 + (.785)(\text{EOG}) + (-.715)(.71) + (-.083)(.71)(\text{EOG})$$

$$= 58.5 + .726(\text{EOG})$$

while a White student's looks like this:

$$\text{Mean EOC} = 59 + (.785)(\text{EOG}) + (-.715)(-.29) + (-.083)(-.29)(\text{EOG})$$

$$= 59.2 + .809(\text{EOG})$$

when EOG is the student's mean EOG score. The contrast between Blacks and Whites is quite similar to that between low SES and other students. Lest one concludes that Blacks and low SES are the same students, within the analytical sample the correlation between the two variables is .42, a moderate but not strong correlation. When EOG scores are average, Whites are predicted to have a mean EOC score nearly a full point above Blacks. With each additional point on the mean EOG score, whites add .8 points to their mean EOC while Blacks add .7 points, on average and net other factors.

The second model in Table 10 adds tracking to the mix. Much of the literature discussed previously states that tracking is responsible for different learning outcomes for races and classes because minorities and the poor tend to enroll in lower-track classes. Adding the proportion honors variable, we see that the more honors classes a student takes, the higher she will score on the EOCs, even when EOG scores (and other factors) are held at their means. Going from no honors classes to all honors classes adds more than 3 points to a student's score, net other factors. Notice also that the estimates for low SES and Black as well as the terms for their interactions with mean EOG scores retain their signs and significance, but each lessens in degree. In other words, different track placements are accounting for some of the gaps between Blacks and Whites and (even more so) between social classes, but not all. For example, when tracking is added to the model, the coefficient for low SES (with EOG scores at their mean) drops from .83 to .6 and the coefficient for Black drops from .715 to .697.

College Achievement

Where we would expect inequality to be more apparent is in the transition from high school to college. Both MMI and EMI theory predict that at levels of education that are not accessible to all, student background will have a greater effect. In the terminology of the theories, the degree to which a level of education is accessible to all groups in society is called its "saturation." At the college level, saturation is not likely to be as high as saturation at the high school level, largely due to the many obstacles to entering and completing college. Among the most obvious are tuition, lost income, and the need for guidance or encouragement. The next section examines the connection between EOC tests and college performance. Table 11 shows that achievement gaps continue to appear at the high school-to-college transition related to student background persist in college.

All three models in Table 11 are cross-classified multilevel models that account for effects of high schools, effects of college campus, and field of declared major (by dummy variables). The first column of the table above shows that background factors significantly influence college GPA. Whites show an advantage over other non-Asian races and males do better than females, as shown in the second model with the addition of the variable for mean EOC scores.

Table 11. Predicting college success with background and EOC scores.

	Model 1		**Model 2**		**Model 3**	
	Estimate	**S.E.**	**Estimate**	**SE**	**Estimate**	**SE**
Mean EOC			.042*	.001	.044*	.001
Low SES	-.170*	.017	-.108*	.016	-.078*	.019
EOC * LowSES					-.009*	.003
Black	-.483*	.014	-.209*	.015	-.173*	.019
EOC * Black					-.010*	.002
Hispanic	-.181*	.050	-.119*	.048	-.105*	.048
Asian	.082*	.034	.111*	.032	.118*	.032
American Indian	-.330*	.054	-.181*	.051	-.162*	.051
Other race	-.197*	.096	-.169	.093	-.167	.092
Male	-.266*	.011	-.310*	.011	-.313*	.011
Limited English	.095	.065	.205*	.063	.212*	.063
STEM major	.597*	.015	.446*	.015	.440*	.015
Professional major	.397*	.032	.451*	.031	.455*	.031
Health major	.683*	.025	.675*	.024	.679*	.024
Humanities major	.573*	.017	.534*	.017	.534*	.017
Education major	.653*	.018	.653*	.017	.655*	.017
Social science major	.526*	.015	.477*	.014	.473*	.014
Constant	2.56*	.006	2.269*	.009	2.254*	.010
Log-likelihood	42358.5		40802.7		40777.8	
N	18647		18647		18647	

** Significant at .05*

The second column shows that when EOC scores are included, the model fit improves but most gaps persist. The estimates for class and race (other than Asian) shrink, suggesting that one of the reasons they were falling behind their counterparts was that they were taking lower-track classes. The male-female gap, on the other hand, increases when course tracks are held constant, suggesting that females earn a higher GPA in college than males, net other factors, and when held to the same level of honors classes, female outperform males even more.

The third column adds interaction terms and shows that, yet again, the benefits that Blacks and low-SES students will gain from improvements at the high school level will not lead to improvements as large as other students, net other factors, when they get to college. The equation using SES, mean EOC scores, and their interaction term looks like this:

$$\text{College GPA} =$$

$$2.254 + (-.078)(\text{LowSES}) + (.044)(\text{EOC}) + (-.009)(\text{LowSES})(\text{EOC})$$

and the equation for Blacks (again, compared to Whites) looks like this:

$$\text{College GPA} =$$

$$2.254 + (-.173)(\text{Black}) + (.044)(\text{EOC}) + (-.010)(\text{Black})(\text{EOC})$$

Entering values for low SES students we get

$$\text{College GPA} = 2.254 + (-.078)(.7) + (.044)(\text{EOC}) + (-.009)(.7)(\text{EOC})$$

$$= 2.199 + .038(\text{EOC})$$

And for other students we get

$$\text{College GPA} = 2.254 + (-.078)(-.3) + (.044)(\text{EOC}) + (-.009)(-.3)(\text{EOC})$$

$$= 2.277 + .047(\text{EOC})$$

As in the middle school to high school transition, we see that low SES students fall behind others with the same level of earlier achievement. When mean EOC scores are average, low SES students earn college GPAs nearly a tenth of a point lower than their peers' GPAs, and as they earn scores above the mean, low SES students exhibit smaller improvements to their GPAs.

Entering values into equation for race gives us

$$\text{College GPA} = 2.254 + (-.173)(.71) + (.044)(\text{EOC}) + (-.010)(.71)(\text{EOC})$$

$$= 2.13 + .037(\text{EOC})$$

for Blacks while for Whites we get

$$\text{College GPA} = 2.254 + (-.173)(-.29) + (.044)(\text{EOC}) + (-.010)(-.29)(\text{EOC})$$

$$= 2.304 + .047(\text{EOC})$$

Again, Whites do better than Blacks both in terms of the intercept (when mean EOC is at the mean) and the slope (which indicates Whites do better than Blacks in college GPA for each additional point scored on the mean EOCs).

Comparing High-Stakes Tests to Other Available Measures

A lot of literature has been devoted to the evaluation of the SAT as a tool for predicting college success. In general, research has demonstrated that a student's score on the SAT is correlated with his or her college grade point average, but its predictive power pales in comparison to high school grade point average (Cohn, Cohn, Balch, & Bradley Jr., 2004). The attraction of the SAT is found in its standardization – people feel that it is "more fair" than grades because all test takers are evaluated on the same criteria, unlike high school grades, which vary in their meaning from state to state, school to school, teacher to teacher, and class to class. EMI theory, however, raises doubt that efforts to make education “more fair” can ever be successful. The SAT may very well provide a measure of academic talent that is independent of a student’s other education credentials, and thus it allows a student from an impoverished high school to be compared to one from an elite private school. But at the same time, the family of the private school student will use its advantages to gain an edge in preparation for the SAT.

The arrival of achievement tests offers another standardized tool for evaluating students and, perhaps, assessing students’ potential for higher education. Some research has shown that achievement tests, which differ from the SAT in that they are specifically designed to reflect course content, show promise in this capacity. It should be acknowledged that achievement tests are not specifically designed to assess college readiness, whereas the SAT is. Of course, high school GPA is not intended to assess college readiness, yet it has time and again proven to be quite useful for that purpose, and consistently more so than SAT.

If achievement tests like the EOC prove to be useful in predicting college success an obvious benefit will accrue to all those involved in college admissions who want convenient, fair, and accurate assessments of student potential. But the greater benefit would be to students themselves. In the era of high-stakes testing, students are under intense pressure to score well on these tests. If it can be established that doing well on EOCs will translate into future educational gains, then investment by schools, teachers, and students in boosting scores will pay off. If, on the other hand, EOC scores do little to predict college performance, then all the effort devoted to raising scores may be little more than a distraction. Table 12 presents the results of these models. Gaps associated with race and social class persist in all models, and we see that high school GPA is better than SAT or EOC scores for predicting college GPA.

Table 12. Multi-level models predicting college academic success with high school GPA, total SAT score, and mean EOC score.

	Model 1		**Model 2**		**Model 3**		**Model 4**		**Model 5**		**Model 6**	
	Est.	**S.E.**	**Est.**	**S.E.**	**Est.**	**S.E.**	**Est.**	**S.E.**	**Est.**	**S.E.**	**Est.**	**S.E.**
High school GPA			.006*	.000					.005*	.000	.005*	.000
SAT Total					.001*	.000			.000*	.000		
Mean EOC							.037*	.001			.008*	.001
Low SES	-.131*	.017	-.104*	.016	-.073*	.016	-.097*	.016	-.088*	.016	-.099*	.016
Black	-.396*	.014	-.141*	.014	-.187*	.016	-.202*	.015	-.098*	.015	-.120*	.015
Hispanic	-.167*	.049	-.140*	.046	-.096*	.048	-.120*	.048	-.119*	.046	-.132*	.046
Asian	.063	.033	.044	.031	.068*	.032	.098*	.032	.047	.031	.052	.031
American Indian	-.324*	.053	-.189*	.049	-.154*	.052	-.196*	.051	-.148*	.049	-.173*	.049
Other	-.147	.095	-.002	.088	-.002	.088	-.015	.092	-.021	.088	-.013	.088
Male	-.260*	.011	-.174*	.011	-.305*	.011	-.030*	.011	-.196*	.011	-.189*	.011
Limited English	.160*	.064	.168*	.059	.239*	.063	.221*	.062	.192*	.059	.180*	.059
Proportion honors	.449*	.017	-.074*	.019	.208*	.019	.207*	.019	-.103*	.019	-.085*	.019
STEM major	.532*	.015	.339*	.015	.443*	.015	.434*	.015	.329*	.015	.333*	.015
Professional major	.418*	.032	.454*	.029	.470*	.031	.454*	.031	.467*	.029	.459*	.029
Health major	.671*	.024	.593*	.023	.684*	.024	.671*	.024	.604*	.023	.598*	.023
Humanities major	.545*	.017	.486*	.016	.504*	.017	.526*	.017	.478*	.016	.486*	.016
Education major	.645*	.017	.580*	.016	.656*	.017	.649*	.017	.589*	.016	.586*	.016
Social science major	.504*	.015	.431*	.014	.462*	.014	.472*	.014	.424*	.014	.429*	.014
Constant	2.453*	.007	2.603*	.007	2.534*	.008	2.255*	.009	2.615*	.008	2.55*	.012
Log Likelihood	41712.6		39116.5		40784.6		40684.2		39051.1		39095.1	
N	18647		18647		18647		18647		18647		18647	

** Significant at .05*

Model 1 presents background variables predicting college GPA without high school GPA, SAT scores, or EOC scores. The coefficients for low SES, Black, Hispanic, and American Indian are each higher in that model than in any of the following models, indicating that a portion of the gaps in their college performance is explained by their performance in high school, but not completely. Models 2, 3, and 4 add high school GPA, total SAT score, or mean EOC, respectively. Each has a positive and significant relationship with college GPA, but the model with high school GPA has the best fit. Notice that when high school GPA is included, the coefficient for proportion honors classes turns negative, suggesting that when high school grades are held constant, an honors student actually does worse than a non-honors student. Why might this be the case? If we keep in mind that honors classes typically receive more credit than a non-honors class, reason becomes clear. A student with a 3.0 high school GPA in a full slate of honors classes actually earned lower grades (Cs that count as Bs) than a student with a 3.0 high school GPA with no honors classes (and Bs that count as Bs). Once they arrive in college, the honors student performs lower than the non-honors student who earned the same GPA, but better course grades. Defenders of academic tracking are often quick to point out that honors classes are more difficult, and therefore the quality point increase is warranted. If that were the case, the adjusted GPAs should be accurate, yet we find that the honors student is not as well prepared as a non-honors student who earned the same GPA without the honor credit boost.

Model 5 pairs high school GPA with total SAT score and model 6 pairs it with mean EOC score. I rejected models with both SAT and EOC because the two variables are so correlated that collinearity became problematic. Comparing model 6 to model 4, a substantial decline in the estimate for mean EOC score is apparent, while the decline for high school GPA is modest. This change suggests that the predictive power of EOC scores is modest compared to high school grades.

The above results suggest that EOC tests, like SAT scores, lose predictive power when high school GPA is included in the model. EMI theory might account for the failure of EOC scores to be better predictors if more privileged students are able to maintain advantages, such as through academic tracking, under accountability reforms. Tables 13 and 14 explore this issue. The former separates upper-track students, defined as those taking all of their EOC courses as honors courses (model 1), from lower-track students, defined as those taking none of their EOC courses as honors courses (model 3). Students taking some combination honors- and non-honors-level EOC courses appear in model 2. Keep in mind that each model includes completely different students from the other models.

Recall that model 6 of Table 12 showed a significant coefficient for mean EOC scores of .008. In model 1 of Table 13, which considers only all-honors students in an otherwise identical model, the coefficient for mean EOC score remains significant and climbs to .020. Mean EOC scores do a better job predicting college performance among upper-track students than they do among students generally. Among students taking a mix of honors and non-honors classes (model 2), the coefficient for mean EOC score is smaller but remains significant. When the model includes only the lower-track students (model 3), the coefficient shrinks further and loses significance, suggesting that EOC scores are poor predictors of college success for students who take regular-track classes.

Table 13. Predicting college GPA for high-, middle-, and low-track students. While high school GPA is a fairly consistent predictor of college GPA across academic tracks, mean EOC is a much better predictor among all honors students than it is among students with only some or no honors courses.

	Model 1: Students with all honors courses		**Model 2: Students with some honors courses**		**Model 3: Students with no honors courses**	
	Estimate	**S.E.**	**Estimate**	**S.E.**	**Estimate**	**S.E.**
Mean EOC score	.020*	.004	.009*	.001	.002	.002
High school GPA	.004*	.000	.005*	.000	.005*	.000
Low SES	-.043	.085	-.107*	.018	-.020	.027
Black	-.142*	.063	-.138*	.016	-.129*	.030
Hispanic	-.174	.145	-.160*	.050	-.053	.103
Asian	-.138	.083	.043	.033	.031	.085
American Indian	-.084	.186	-.129*	.052	-.197	.117
Other race	.130	.288	.059	.100	-.216	.168
Male	-.320*	.037	-.226*	.012	-.126*	.024
Limited English	.145	.233	.164*	.068	.066	.111
STEM major	.401*	.049	.492*	.016	.478*	.044
Health major	.530*	.083	.748*	.025	.695*	.039
Business major	.534*	.054	.674*	.018	.650*	.036
Humanities major	.481*	.053	.617*	.018	.695*	.039
Education major	.485*	.061	.730*	.018	.778*	.036
Social science major	.318*	.051	.532*	.016	.719*	.030
Constant	2.135*	.205	2.492*	.012	2.622*	.024
Log Likelihood	2789.2		28819.0		7379.1	
N	1406		14339		3595	

** Significant at .05*

While Table 13 reveals that the way mean EOC scores predict college GPA depends in part on a student’s academic track, Table 14 explores this relationship further by using an interaction term rather than separating students into different models by academic track. The interaction term in this case is the product of the proportion honors classes and mean EOC score variables.

Table 14. Predicting college GPA with an EOC score and proportion honors interaction. The more honors courses a student takes, the greater the benefit EOC performance has for college GPA.

	Effects on college GPA	
	Estimate	S.E.
Mean EOC scale score	.035*	.001
Proportion honors	.135*	.028
EOC * Honors	.010*	.003
Low SES	-.100*	.016
Black	-.209*	.015
Hispanic	-.120*	.048
Asian	.098*	.032
American Indian	-.194*	.051
Other race	-.149	.092
Male	-.303*	.011
Limited English	.221*	.062
STEM major	.432*	.015
Professional field major	.456*	.031
Health major	.672*	.024
Humanities major	.526*	.017
Education major	.649*	.017
Social science major	.471*	.014
Constant	2.262*	.009
Log Likelihood	40682.0	
N	18647	

** Significant at .05*

Coefficients for percent of classes that are honors level and mean EOC scores are positive and significant. When students score at the grand mean on their mean EOC scores, their college GPA will be higher the more honors courses they take in high school, net other factors. The interaction term that is the product of those two independent variables is positive and significant, indicating that higher EOC scores do indeed predict higher college GPA, *especially if the student takes higher track classes*. Put another way, high EOC scores are more likely to lead to a strong college GPA for honors students than for others.

Based on the figures in Table 14, the explanatory and interaction terms form an equation that looks like this:

$$\text{College GPA} = 2.262 + (.035)(\text{EOC}) + (.135)(\text{Honors}) + (.010)(\text{EOC})(\text{Honors})$$

The minimum and maximum of proportion honors in the analytic sample is .2 and .8, respectively. An equation for an all-honors student would look like this:

$$\text{College GPA} = 2.262 + (.035)(\text{EOC}) + (.135)(.8) + (.010)(\text{EOC})(.8)$$

$$= 2.37 + (.043)(\text{EOC})$$

And for a no-honors student would look like this:

$$\text{College GPA} = 2.262 + (.035)(\text{EOC}) + (.135)(-.2) + (.010)(\text{EOC})(-.2)$$

$$= 2.24 + (.033)(\text{EOC})$$

Students who take all of their EOC courses at the honors level will boost their college GPA by a greater amount with an improvement in their mean EOC score than a student with lower-track classes making the same improvement.

Tests of Assumptions

I conducted several tests of assumptions of multiple regression on each of these models. First, I tested for multicollinearity using the variable inflation factor (VIF) calculation included in software packages. A high level of inflation (say, 5 or more) can lead to larger standard errors. The highest VIF in most models presented in this chapter was about 2.3, and it occurred when high school GPA and mean EOC scale scores appeared in the same models.

I also tested the distribution of residuals for normality. The models that included EOG and EOG scores as dependent variables cleared this test without any difficulty. The models in which college GPA was the dependent variable were nearly perfectly normal, but showed a slight skew. The source of the skew appears to be in the presence of outliers in the models that include high school predictor variables and college outcome variables. Specifically, there are a handful of students who scored particularly low high school GPAs yet earned high college GPAs. Although high school and college grades are highly correlated, it is understandable that occasionally students will struggle at one level of education and thrive at the next. Consider the gifted underachiever who is bored in high school and inspired in college, the teenage misfit who needs college as an escape from a miserable home and social scene, the student who gets sick in high school or whose parents split up and misses a lot of classes in high school but manages just fine in college, or the student who attends a highly competitive high school (and takes non-honors classes) and then attends a relatively easy college. In any of these cases, a student could have a poor high school GPA yet earn high marks in college.

Dealing with outliers is tricky business, especially when there is no theoretical reason for eliminating them from the study. In such situations, it may be best to run the tests both with and without the outliers (Pedhazur & Schmelkin, 1991). Excluding the outliers produced results that were substantively similar to those that came from models that included the observations, but included estimates for the EOC coefficients that were more drastic. Put another way, dropping the outliers resulted in more compelling results, at least in terms of my hypotheses. However, I chose to present the more conservative findings in this results section.

The next chapter analyzes the results presented above more thoroughly.

CHAPTER 5: DISCUSSION

This dissertation examines the roles that high-stakes testing accountability programs and academic tracking play in the reproduction of social inequality. The previous chapters presented the research questions, a review of the literature, a description of the methods, and a presentation of the results. Previous research indicates that achievement gaps along lines of race and social class are common at all levels of education. It also shows that high-stakes testing and tracking contribute to the achievement gap by pressuring teachers of low-achieving students to spend more time on test preparation at the expense of higher-order thinking skills. The results of the multi-level regression models used with data from one complete cohort of North Carolina public school students seems to support earlier findings. Results also support the notion that high-stakes testing exacerbates the negative consequences of academic tracking. This chapter discusses those findings.

The research questions identified in the second chapter asked whether the outcomes of North Carolina's high-stakes tests reflect achievement gaps along the lines of race and class, whether those gaps seem to get larger or smaller over students' educational careers, and whether tracking alters the effects of high-stakes tests. The first two hypotheses proposed that middle and high school tests will reflect achievement gaps, even when controlling for prior achievement. The third through fifth hypotheses state that those gaps continue in college, academic tracking contributes to those gaps, and high-stakes tests are poor predictors of college performance, particularly for lower-track students.

High School Achievement

The results in Table 9 indicate that the students in North Carolina exhibit significant achievement gaps along lines of race and class, as measured by the state's required standardized tests. In terms of students' average performance on middle school EOG tests, males do worse than females; Blacks, Hispanics, and American Indians do worse than Whites; and Asian students outperform all other race/ethnic groups. Students with limited English skills do worse than other students and students of low socio-economic status do worse than other students. Considering previous research on achievement gaps, none of these results is surprising.

The second model in the table uses mean EOC scores as the dependent variable. EOC tests resemble EOG tests in format, purpose, context, standard deviation, and range. The means are different, but otherwise these measures are quite similar. The difference between the means of each dependent variable is responsible for the substantial shift in the intercept between the two models. The coefficients, however, remain quite similar. Compared to the middle school scores, on EOC tests Blacks, Hispanics, American Indians, low SES students, and limited English students closed the gaps slightly with their counterparts. Asians extended their advantage over Whites.

The third column of Table 9 includes EOG results used as a measure of prior achievement in a model predicting EOC results. We might expect that a student's performance in middle school will be roughly equivalent to her performance in high school. Indeed, the results indicate that each point a student scores on the EOG is associated with nearly eight-tenths of a point on her average EOC score, net of other factors. The other independent variables in that same model indicate whether gaps increased or decreased as this cohort moved from middle school through high school while holding EOG scores constant. Interpreting the dummy variables is tricky because the variables are centered on the grand means. To see the effect of SES, we multiply the estimate by .7 for low SES students and by -.3 for other students. The result for low SES students is (.7)(-.666)=-.466 and for other students is (-.3)(-.666)=.2, meaning that low SES students score, on average, .2-(-.466) = -.666 below other students (the same as the estimate). The gap between Blacks and Whites grew. Relative to Whites, Blacks on average scored .441 points lower than their prior achievement would have predicted. American Indians fell even further behind. These differences sound substantively small, but they are statistically significant.

Hispanics, as a group, did better than their prior achievement would suggest, relative to Whites. A likely explanation for Hispanics closing the gap while Black and Native American students fell further behind could be found in the fact the same students are being tested at two different times in their educational careers here in North Carolina, and the improvement in scores going from 7th to 12th grade may reflect the success of immigrant and Spanish-speaking individuals among the Hispanic population[9] to assimilate into or accommodate the dominant culture (see Gibson, 1988). Asian students also did better in high school than their middle school performance predicted, perhaps for the same reason: the immigrant students among them made gains that resulted from

[9] The U.S. Census Bureau reports that more than 6% of North Carolina residents speak Spanish at home.

improved assimilation or accommodation into schools and communities.[10] That explanation is supported by the limited English variable. While these variables did not violate assumptions about collinearity, they followed similar patterns, perhaps because not all students for whom English is a second language would be labeled “Limited English.” Students in that category performed more than a full point better than their prior achievement predicted, relative to other students and net other factors. Improvements in their ability to read, write, speak, and comprehend English in the years covering middle and high school offer a reasonable explanation.

Overall, the results of Table 9 tell us what prior research and theory had predicted: achievement gaps associated with race and social class persist. Additionally, we see support for the first two hypotheses. First, Whites perform better than Blacks, Hispanics, and Native Americans in both the middle school and high school standardized tests. Low SES students lag behind their counterparts in both tests as well. The second hypothesis predicted that gaps would exist even when controlling for prior achievement, because privileged families find ways to maintain their advantages, even in nearly-universal levels of education. While Asians and Hispanic students did better on their EOC tests than their EOG scores would have predicted, Blacks and American Indians slipped, relative to Whites, and low SES students slipped compared to others as well.

Table 10 tests this notion further. In the first column, interaction terms tell us that race and class moderate the effects of prior achievement on mean EOG scores. The calculations performed in the previous chapter showed that at the sample’s mean EOG score, low SES students score nearly a point lower than other students, and then for each

[10] The Migration Policy Institute reports that nearly a quarter of foreign-born North Carolina residents migrated from Asia.

point scored above the mean on the EOGs, low SES students gained about a tenth of a point less than other students gained. Similarly, Blacks score nearly a point lower than Whites on EOC tests when they score at the mean on EOG tests, and then gain about a tenth of point less than Whites for each additional point scored on the EOGs. These differences are small, substantively speaking. Even at the extremes, the gaps associated with class and race come nowhere close to a full standard deviation. That these differences are small may be to some degree reassuring, but that does not make them inconsequential. It is one thing for a public education to try to close the gap that exists when children first arrive in schools. It is another for that gap to grow wider after those children have been in school for several years.

The models used in this dissertation are multi-level models. Students are clustered in schools, and schools differ in their resources, teaching quality, parental support and involvement, and need to prepare for high-stakes tests. These differences are fascinating and rich topics for research, but beyond the scope of this dissertation. Moreover, a lot of research has already been done about the effects of high-stakes tests on different schools. Instead, this dissertation explores another source of education differences in the form of academic tracking. The second model in Table 10 adds the proportion honors variable to the others. The estimate is positive and significant, indicating that going from no honors classes to all honors classes is associated with a jump of more than three points on mean EOC scores, even when controlling for prior achievement and background factors.

Most of the other estimates shift only slightly between the two models. The values that shrink, such as those for Blacks, Hispanics, Asians, and SES, do so because differences in academic tracks account for some of those differences in EOC scores. The interaction terms maintain their basic relationships but shrink slightly as well. Interestingly, the estimate for American Indians increases in size, indicating that the gap between this group and others gets even worse when tracks are held constant.

Table 10 reveals three key relationships. First, it shows the moderating effect that race and class have on achievement, giving support to the hypothesis that achievement gaps get worse, not better, as students progress through high school. Second, it reveals that tracking plays a major role in student outcomes. The difference in outcomes associated with tracking is greater than the race, class, language, and gender variables included in the second model in Table 10. Third, the gaps associated with race and class persist and are not explained away by either prior achievement or tracking. Prior achievement and tracking do account for some of the differences, but not all. I explore the issue of tracking further later on. Before that, however, I introduce a different dependent variable, college GPA.

College Achievement

More than 20,000 students in North Carolina’s public school Class of 2004 entered the UNC system after high school. Table 11 presents three models that predict students’ college GPA. Model 1 includes background variables, model 2 adds EOC scores, and model 3 adds interaction terms. All three models, as well as those going forward, are cross-classified, three-level multiple regression models that allow for the grouping of students in high schools and in college campuses. The models also include six variables to control for different majors.

Model 1 shows that gaps associated with race and class continue in college. Model 2 shows that gaps get somewhat smaller when previous achievement, as measured by mean EOC scores, are added to the model. Differences in prior achievement account for a portion of the gap between classes, a portion of the Black-White gap, a portion of the Hispanic-White gap, and a good deal of the American Indian-White gap. The advantage Asians have over Whites in college GPA gets even greater once EOC tests are added to the predictor variables. Females also extend their lead over males who had similar past achievement.

If we assume that mean EOC scores are a valid measure of high school achievement,[11] we see a few interesting things taking place as students make the transition into and move through college. Holding EOC tests constant, low SES students earn a college GPA a tenth of a point below other students, Blacks earn a college GPA two tenths of a point below Whites, and Hispanics and American Indians earn GPAs a little more than a tenth of a point below Whites.

The third column of Table 11 adds interaction terms to the model. Again, these results show that race and class moderate the effects of EOC performance on college performance in small but statistically significant ways. As we saw in the progression from middle school to high school, as students go from high school to college, those from disadvantaged groups fall behind other students even when previous achievement is held constant. Considering the estimates for race and class in all three models, Table 11 offers fairly persuasive evidence confirming the third hypothesis, regarding student background as a predictor of college achievement.

EMI theory explains all of these findings. Parents of means are able to provide their children with not only quantitatively more education, but also qualitatively better

[11] The issue of validity among high-stakes tests has a literature all its own. There are significant reasons for skepticism regarding the validity of such tests. The conclusion of this dissertation identifies some of them.

education. Hiring tutors, purchasing educational texts and software, providing sufficient times and locations for homework are some of the likely advantages that privileged parents secure for their children. As students move through middle school and into college, these parents can help with chemistry and calculus homework because many of them similar courses in the past. Moreover, privileged parents model the kinds of attitudes and behaviors needed to focus on homework and studying when students are not in school (Lareau, 2003). Rather than fade away as children become more independent, these gaps seem to build on each other as students continue through school. As Lucas himself points out, academic tracking can be a mechanism by which the gaps continue to grow over years of education.

So far, the examination of factors contributing to college success has relied upon EOC test scores as the measure of high school achievement. Before I get into the effects of tracking as well, I will compare mean EOC test scores to other measures of high school success, namely SAT and high school GPA.

The Predictive Power of SATs and High School GPA

So far, I have found evidence that (1) achievement gaps associated with student background in North Carolina do exist, (2) those gaps grow over the educational careers of students, and (3) improving one’s middle or high school performance yields greater dividends for privileged students than for Black, American Indian, or low-SES students. The next step is to evaluate whether high-stakes tests are the best way to predict future achievement.

The results appearing in Table 12 include 6 models, each with controls for student background, tracking, and school effects, college major, and a different combination of

EOC scores, SAT scores, and high school GPA. The first model is simply a baseline assessment that gauges the effects of background and tracking on college GPA without any controls for prior performance. It is very similar to the first model in the previous table, except that it includes a control variable for tracking. Not surprisingly, the tracking variable is positive and significant – the more honors classes a student takes in high school, net other factors, the higher that student's college GPA is likely to be. Even with this control variable, however, gaps exist along lines of race and class.

The second model of Table 12 adds what the literature tells us is consistently the best predictor of college grade point average: high school grade point average (Cohn et al., 2004). The causal mechanism here is easy to imagine – the same characteristics that cause someone to get a good GPA in high school probably carry over into college. The effect of tracking shifts drastically once GPA is added, presumably because (1) students taking difficult courses in high school also take difficult courses in college and (2) high school GPA is heavily weighted, typically a full point, in favor of honors, AP, and IB classes. If two students on different academic tracks achieve the same high school GPA, having taken honors classes means earning lower letter grades because the GPA includes weights boosting the value of grades in honors classes. While the addition of high school GPA weakens the race- and gender-based coefficients, it actually increases the limited English coefficient, which is positive in both models, indicating that the gap in college performance associated with language differences increases when high school GPA is held constant. Why would a limited English student do better than a proficient speaker? One reason may be that language is less of a barrier to navigating college than it is for high school. Another reason is that by the time these students got to college, they had improved their English proficiency.

The third model of Table 12 uses total SAT score as the sole academic predictor. It does not predict as well as high school GPA, as evidenced by the larger coefficients for males and Blacks and the larger log likelihood. Differences associated with SAT score account for less than half of the Black-White gaps and the SES gap, judging by the shifts in their coefficients. The fourth model is more relevant to our discussion of high-stakes testing. In it, average EOC score is used to predict college GPA. This model has a poorer fit than either the high school GPA or the SAT models, but the background variables, overall, are not much different from the other two models. So far, EOC scores seem to be similar to high school GPA and SAT scores in terms of their ability to predict college performance, net other factors.

Models 5 and 6 consider combinations of academic indictors in the prediction of college GPA. Model 5 indicates that using high school GPA and SAT together, a common practice of college admissions officers, improves the model beyond using either of those predictors alone. But while model fit is improved, the gaps associated with student background persist at levels similar to the other models. One exception is the estimate for Blacks, which drops dramatically when both high school GPA and SAT total score are included in the model. Model 6 replaces SAT scores with EOCs. This switch has implications for the debate over achievement tests, which measure whether a student has learned something, vs. aptitude tests, which measure whether a student has the potential to learn something. In the case of these two models, combining high school

GPA with SAT scores evens out the student background effects slightly better than combining high school GPA with mean EOC, but the difference is small.

One thing that makes EOC tests stand clearly apart from both high school GPA and SAT scores is the issue brought up by opponents of accountability programs: EOC tests are high-stakes for teachers and schools. All three assessments are high stakes for students, and the distance that some students will go to get high marks on any of them is both familiar and remarkable. Teachers' and schools' accountability evaluations, however, do not depend on SAT scores and grades, at least not formally or to the degree that they depend on EOC scores. Every teacher, principal, and superintendent, not to mention student, has incentives to try to game the system and get higher EOC scores for their students than the students' actual abilities would likely produce.

What light might Lucas's EMI theory shed on this question? If we start with the premise that all students want high grade point averages, SAT scores, and EOCs, EMI tells us that along the way to securing those measures of achievement some students will have access to more education opportunities than others. In other words, two students may leave high school with a 3.8 GPA, but one may actually have received a qualitatively better education by virtue of learning things outside the scope of the grade point average measures. For example, while high school teachers often reward effort by giving generous points for homework completion or participation, college professors emphasize product. While the SATs emphasize the ability to pick out the correct answer from among choices on a pen-and-paper (or, increasingly, a computerized) test, college courses require sustained persistence, communication skills, and presentation abilities. EOC tests have similar shortcomings in terms of their abilities to cover all the aspects of learning that will reflect the potential to succeed in college. To compare EOCs, SATs, and high school GPAs head-to-head I put all three in a model similar to those in Table 12. The results contained high levels of collinearity between SAT and EOC scores, so I left it out of the results chapter. What was telling about the results, however, was that while coefficients for high school GPA and SAT resembled those in the other models, the sign for mean EOC became negative. Although methodologically suspect, the results suggest that EOCs are not nearly as good at predicting college GPA as the other measures. Why might that be the case?

Recall that earlier tables showed academic tracking to have a profound effect on academic achievement, both in terms of predicting EOC scores and college grade point average. Recall as well that Lucas found academic tracking to be a good example of EMI at work, because it allows privileged families an opportunity to separate their children from others and get them into classes with more challenging work and more valuable learning. A higher track means a higher GPA, partly because better students sign up for higher tracks, partly because they learn more, and partly because they get a boost to their GPA built in to the calculations. A higher track means a higher SAT score, largely for the first two reasons mentioned above. Presumably, a higher track also means a higher EOC score, again for those two reasons. But EOC scores take on a different meaning when it comes to tracked classes. While all students (or nearly all) want high GPAs, SATs, and EOCs, their teachers are mostly concerned about the EOCs. Their own evaluations depend on them, their administrators are worried about them, and their schools may be sanctioned for them.

A complex situation exists where tracking, high-stakes testing, and learning intersect. Middle class parents have seen to it that their students are sorted into the upper tracks. Teachers trust that the upper-track students will be proficient on the high-stakes exams and are more concerned that their lowest-ability (and thus lowest-tracked) students will do poorly. As a result, the privileged students, safe in their honors, AP, and IB classes, are not exposed to the emphasis on preparation for the EOC tests. In contrast, the students relegated to the lower-track classes are bombarded with pressures and lessons emphasizing readiness for high-stakes tests. While everyone learns with an eye on the EOC tests, the higher-track classes also enjoy opportunities to learn far more. Their teachers and parents expect them to pass the EOCs and expect them to do well in college. As a result, they get an education that balances those two expectations. For lower track class, parents and teachers expect most to pass the EOCs and some to go to college. Therefore, they emphasize the academic tests that they are certain their students will face – the EOCs – and de-emphasize the academic tests that a few will face – college. In other words, EOCs and all high-stakes tests distract teachers and students in low-track classes from learning and teaching the lessons that will best prepare them for the next level of education. In the end, the low-track students who are especially motivated and bright will score very well on their EOC tests, but when they get to college, they will find it difficult to compete with the students who took honors classes and perhaps scored the same as they did.

The Role of Tracking

The results in Tables 12 and 13 serve to test this notion that track placement will determine whether the connection between EOC scores and college GPA is influenced by the track a student takes. We expect that higher-track students will score better on both EOCs and college GPA, but what we want to know is whether tracking distorts the ability of EOC scores to predict college GPA. Tables 12 and 13 investigate this question in two different ways. In the former, low-, mixed-, and high-tracked students are considered separately, and with different outcomes. In it we find that students taking all honors courses display a positive and significant relationship between their mean EOC score and their college GPA, a relationship that is much stronger than the one found in a similar model that included all students. The relationship is weaker for the mixed-track students, and even more so for the non-honors students. Those results suggest that indeed, honors students are learning college-preparation lessons alongside their test preparation, while the lower-track classes might be focusing on the lessons that prepare students for the exam but not for further education.

Table 14 adds an interaction term between EOC score and tracking, and considers all students in one model. According to its results, EOC scores predict college GPA more accurately for students with more high-track classes. At the grand mean EOC score, students in all honors classes score a tenth of a point higher than students in no honors classes, and for every additional point they earn in their mean EOC score, the all-honors class students add 1/100th of a point more than no-honors students to their college GPA. The standard deviation for mean EOC score is about 8, so at two standard deviations above the mean EOC score, all-honors students are about .26 points higher in their college GPA than a no-honors student, or about a third of a standard deviation.

Many would argue that this difference is too small to be worthy of a policy change, yet such results are empirical evidence of the very outcomes the critics of high-stakes tests warned about. The stakes for test-based accountability policies are especially high for low-achieving students and their teachers and administrators. Tracking concentrates these students in particular classrooms where teachers, as rational actors concerned about assessments of their job performance, narrow the curriculum and engage in lessons that prepare their students for multiple choice tests, the format for all of the North Carolina EOC and EOG tests. The resultant scores may indeed accurately reflect a student's mastery of the key concepts in the course, but they bear little similarity to the degree of success that student will have at the next level of education. In contrast, students enrolled in honors classes are unlikely to fall below "proficient" on their tests. Teachers of those classes are not under as much pressure to prepare for the EOC tests, so they feel free to engage in higher-level thinking exercises that may or may not be useful on the EOCs, but will certainly be useful at the next level of education.

Not Even on the Radar

Policymakers, by and large, are not eager to eliminate tracking practices from schools. A recent panel discussion including some high-profile policymakers addressed equal access to education. The panel included a former governor and supporter of school integration, a school board member for a large urban school district, and the superintendent of the same district. An education professor rounded out the panel. In more than an hour of discussion and questions and answers about equal access to education, not a single one of the participants or audience members mentioned academic tracking.

At the end of the program, I approached each of the participants individually and asked what role, they thought academic tracking might play in perpetuating inequalities in education. The former governor, who minutes earlier had spoken passionately about the moral and educational hazards of segregated schools, insisted that academic tracking does not segregate students. “If the school is diverse,” he said, “the classrooms will likewise be diverse, tracking or no tracking.” The school board member echoed that sentiment and added, “we have to make sure every school offers a range of classes, especially Advanced Placement.”[12] When pressed on the question of whether high-stakes tests might drive teachers to short-change low-tracked students, she argued that the problem lies with the misuse of test data, not tracking. That two politicians refused to criticize tracking policies is hardly surprising, as efforts to “detrack” schools, especially diverse schools, have been met with resistance from families with political and cultural capital (Wells & Serna, 1996).

The superintendent differed from the other two only slightly in that he said tracking, as people usually think about it, can sometimes lock students into a particular trajectory and thereby reproduce inequality. He went on to say that he believes in “grouping and re-grouping,” in which students are placed in classes that offer what they need at that particular point, and then reassessed and reassigned as necessary in a fluid and dynamic process. To illustrate his point, the superintendent asked me not to think of the different course level offerings as “tracks, but more as ladders, by which a student can acquire the skills they need to move up.” Lack of mobility was the problem in his eyes, not differences in how honors and non-honors classes are taught. His argument relies on the assumption that a lower-level class will actually provide a motivated student with the

[12] The school board member’s comment sounds like it came right from the mouths of California education policy makers in the late 1990s when they expanded AP course offerings for low-income students and in small schools, only to find that doing so prompted more privileged schools to increase their AP offerings and thus maintain their advantage (Klugman, 2011).

necessary experiences that will enable them to catch up to the honors students and join them in a future class.

The education professor conceded that tracking can harm equity, but noted that he does not see any practical alternative. To group students at four or five different levels of academic ability into one classroom, he explained, and to think that the teacher will successfully differentiate instruction for every lesson is to believe in “a myth.” Therefore, the professor argued, one cannot leap to the conclusion that all tracking practices are always harmful to educational equity.

Given the above reactions to my questions, I doubt the small estimates that my models exhibit will be terribly persuasive to policy makers. The fact that the gaps I find are modest is only part of the problem. The leaders I interviewed, particularly the elected officials, were doubtful of the causal mechanism I believe exists linking tracking, testing, and achievement. Some qualitative evidence is required to not only verify that teachers do behave differently in different level classes when high-stakes tests are involved but also to illustrate to those in power the nature of these connections. The next chapter discusses such research.

CHAPTER 6: A TALE OF TWO TRACKS

This dissertation explores the effects of high-stakes testing and academic tracking on the reproduction of social inequality. Previous chapters outlined findings in the existing literature, identified the methodological approach, and reported and discussed the findings. As in the literature, this study finds that achievement gaps exist along lines of race and SES. Those gaps tend to increase in magnitude as a cohort of students moves through years of public schooling, a phenomenon predicted by Lucas's theory of Effectively Maintained Inequality. Although there are many mechanisms at work that create and influence the achievement gaps, this study focuses on academic tracking. The results suggest that academic tracking exacerbates achievement gaps because minorities and low-SES students tend to be placed in regular rather than honors classes and those classes offer fewer opportunities to learn. This chapter identifies qualitative evidence from recent studies that support this causal mechanism.

The evidence presented thus far is strictly quantitative in nature. Conclusions regarding the causal mechanisms at work among variables based solely on quantitative evidence can be shaky; qualitative findings, usually obtained through cases studies and interviews, can provide credible support to quantitative research (King, Keohane, & Verba, 1994). Certainly my own experience as a teacher of students in both public and private high schools, high-stakes and no-stakes testing environments, and in both tracked and detracked classrooms have informed the manner in which I conducted this study and analyzed the results.

Indeed, my upper-track classes include ample opportunities to engage in higher-level thinking activities and offer lessons that provide students with opportunities to explore topics in ways that suit their individual curiosities. I have challenged them with Supreme Court role plays, mock trials, public policy research projects, and interviews with elected officials. I have provided my lower-track students those same opportunities, at least for the first eight-years of my career, when I taught high school seniors in subjects for which there were no standardized tests. Upon moving to North Carolina and taking a job teaching 9th and 10th grade social studies, I faced for the first time the pressures of the high-stakes standardized test. Grade 9 was no big deal; it does not have an EOC test in this state. I presented identical lessons in the honors and regular classes for that course and simply held honors students to a higher standard of performance. 10th grade, however, is the year for North Carolina’s Civics and Economics class, the current incarnation of the Economics, Legal, and Political course, and it does have an EOC exam. The two topics, civics and economics, provide wonderful opportunities for engaging lessons. But with the specter of an End-of-Course test looming, I found myself differentiating my teaching between the honors and the regular classes. In my honors class, I’d first cover a set amount of material for the week, make sure the students grasped the content, and then use the remaining time for more challenging learning activities. In my regular classes, I felt I had no such luxury. Because I worried that some students in that class would be at risk of missing proficiency, I skipped a lot of my more engaging lessons and stuck to the basics. In the end, my students, overall, scored well, but

I know I did not give the lower-tracked students opportunities to learn on par with the opportunities I provided to my honors students.

Conversations with colleagues, parents, administrators, and professors have also convinced me that the results I obtained and presented in the preceding chapters reflect a very real mechanism in which upper-track students, shielded from the panic associated with high-stakes tests, receive more thorough and challenging instruction making them better prepared than the their lower-track counterparts for the next level of education. One particularly illuminating discussion was with a former principal of an elementary school in a large, urban North Carolina school system. He described to me in detail the maneuverings that he and other principals would use in children's educations to make sure that certain students would be in class for the tests but not others, that teachers understood which students should get what treatments, and that students at the greatest risk of failing proficiency would have their art and music lessons replaced with remedial drills. "All the principals did it," he remarked.

Many qualitative studies have shown that educators across the country are adjusting instruction to increase pass rates, even at the expense of best teaching practices. Wayne Au's (2007) qualitative metasynthesis of 49 individual studies found that, with only a few exceptions, high-stakes testing environments are associated with the contraction of course content, the fragmentation of knowledge, and teacher-centered pedagogy. Most of the exceptions took the form of social studies or language arts teachers adding new content in preparation for the state tests. Other examples included the inclusion of student-centered learning in the form of social studies students evaluating historical documents, also in preparation for the tests. High-stakes testing deserves some credit for having caused these changes, but only if one assumes that they are replacing inferior content or teaching strategies.

In North Carolina, for example, teachers reported to researchers that the state’s emphasis on getting “back to basics” in its accountability program marginalized art and science education (Jones et al., 1999). In another study Birk (2001) showed that high schools are dropping or paring back lessons on social justice, a topic often requiring higher-level thinking skills such as analysis, synthesis, and evaluation, but rarely getting attention on accountability assessments. Perna and Thomas (2009) studied the effects of high-stakes testing in 15 high schools in five states and found that graduation, academic preparation, knowledge, and information all suffered in the high-stakes testing environment, especially in low-SES schools, making it more difficult for students to learn about and prepare for college.

Investigators also find that the burdens of high-stakes testing are not distributed evenly within schools. Dillon (2006), while writing for the New York Times, visited a high-poverty junior high school and reported that the lowest-performing students take only reading, math, and physical education, while other students enjoy a full slate of courses. In a series of four case studies of urban elementary schools, John Diamond and James Spillane (Diamond & Spillane, 2004) found that educators facing sanctions for low student performance put their focus on the particular students, grades, and subject areas that put their school most at-risk for receiving sanctions. On the other hand, educators whose students earned high marks did not zero in on particular students but distributed their efforts more evenly across the student body. Brown and Clift (2010) visited elementary and middle schools to see how they were responding to accountability policies. When schools faced possible sanctions, they would increase the alignment between curriculum and class content and, in the worst cases, perform educational triage by ignoring the students who were likely to pass or fail no matter the intervention and instead devote their time and resources to those students who were “on the bubble” and could go either way. These findings match those described by Booher-Jennings (2005) in her investigation of Texas schools and the triage conducted by teachers reacting to institutional and peer messages that high test scores equal good teaching.

Of course, most middle and high schools already have low- and high-performing students sorted into classrooms through tracking. The contrasts between high- and low-tracked classrooms in terms of opportunities to learn, instructional methods, academic rigor, and teacher ability was reported decades ago (Oakes, 2005; Page, 1989, 1991). High-stakes accountability reforms, instituted after much of the research on tracking occurred was expected to provide sufficient incentive (and punishment) to both students and teachers in low-performing classrooms to improve their effort and strategies, regardless of track. Indeed, the entire “standards” movement, a key component to accountability, is predicated on the idea that all students should achieve at some minimum level.

Gamoran reports that more recent studies show that tracking continues to magnify social inequality by providing benefits to upper- but not lower-track students. While he acknowledges that lower-track classes can improve student achievement if they include high-quality instruction and high levels of academic rigor, he notes that typically such is not the case (Gamoran, 2009). Milner (2004) conducted a study of a school in Pennsylvania that moved from heterogeneous classes to tracked classes in an effort to improve achievement. Overall achievement did not improve once tracking was put in place, and the student body, which had previously exhibited remarkable solidarity, became fragmented, competitive, and status-conscious. The manner in which the school's academic tracks were conceived and implemented, as well as the new tracks' effects on the social structures within the school, reflected the core of Lucas's EMI theory.

One way in which the practice of tracking has changed since Jennie Oakes' seminal work on the topic (Oakes, 2005),[13] is that students now have more choice. Lucas reported that schools increasingly permit students to select their track course-by-course, rather than committing to an entire slate of either upper- or lower-track classes (Lucas, 1999). Research by Mickelson and Everett (2008) call this new form of differentiation "neotracking", and document not only its use in North Carolina but also the ways in which its implementation so closely resembles tracking in the past.

Such research has not fallen entirely on deaf ears, and several schools and districts have abandoned tracking policies in favor of heterogeneous classrooms and, in at least a few cases, demonstrated that doing so can indeed improve student outcomes for all. In New York state, for example, a study comparing achievement in terms of Regents and International Baccalaureate diplomas before and after tracking reforms found that students in minority and low-SES groups displayed greatly improved outcomes without any decline among groups who had performed well under the previous policies (Burris, Wiley, Welner, & Murphy, 2008). Boaler and Staples (2008) had similar findings in their study of three California High Schools. In that study detracked students made significant gains on state-mandated tests when their new classrooms included the "best practices" of teaching that had previously been associated with only upper-track classes, rather than

[13] Oakes' first edition of *Keeping Track* was published in 1985.

reflecting a “dumbed down” version of the course to which they might previously have been assigned. Tracking is still firmly entrenched in educational policy, institutions, economics, and culture, but evidence is mounting that says heterogeneous classrooms have profound potential to close achievement gaps and raise overall achievement (Gamoran, 2009).

This dissertation focuses on the intersection of these two hotly-debated education policy issues: tracking and high stakes testing. My arguments in the preceding chapters hinge on the idea that instructional responses to accountability policy differ in upper- and lower-track classes within schools. Quantitative analysis appears to support that notion. The studies described above support the first notion that instruction changes in response to high-stakes testing policies, as well as the second, that opportunities to learn continue to differ among academic tracks. Briefly, I argue that high-stakes testing exacerbates the inequalities associated with academic tracking. Maika Watanabe’s (2008) ethnographic case study of a public middle school in North Carolina offers strong support for the mechanism described in my study. I describe her study below.

Watanabe accumulated 68 hours of observation, 12 teacher interviews, and student work samples from two language arts classrooms to learn how teachers differentiate their instruction across academic tracks under high-stakes test policies. She also conducted interviews with eleven other teachers at other schools. The school in question might be considered a “good” public school: it enjoyed strong leadership, parent support, high morale, and an experienced faculty. Although it did not meet its expected growth under NCLB guidelines, most of the students were achieving at or above grade level. Those two characteristics gave the school a “no recognition” label under North

Carolina's classification system. The designation is below "exemplary" and "expected", but superior to "low performing", and carries neither rewards nor sanctions. While the student body overall is evenly split between Black and White, both of the focal teachers in the study had upper-track classes that were majority White (about 80%) and lower-track classes that were majority Black (about 60%). Tracking in the school also divided students by SES: only 5% of the students in upper-track classes received free or reduced lunch compared to nearly one quarter of the students in lower-track classes. These patterns match those described in much of the literature on tracking (Farkas, 2003; Lucas, 1999; Mickelson, 2001; Oakes, 2005).

In her investigation of tracking and accountability, Watanabe was surprised at how similar the curricula were in the upper and lower tracks. The two teachers used the same readings and taught similar skills in both their upper- and lower-track classes. In fact, the teachers apologized to her for repeating the same lessons in each class.[14] However, Watanabe noted several key differences between the two tracks, differences that appeared in both the teachers' lessons and were supported by interviews of teachers at other schools.

The tracked classes Watanabe observed differed in five important ways. First, upper-track classes received less explicit test preparation. While 46% of the instruction in the lower-track classes had a high correspondence to the demands or format of the high-stakes test, only 15% of the instruction in the upper-track classes was highly correlated in the same way. In the place of instruction highly-correlated to the state test, teachers gave upper-track classes creative writing assignments, independent projects, and activities that

[14] It is interesting that the teacher apologized, not only because Watanabe did not expect to be entertained, but because one of the goals of accountability is to make lessons and curricula more consistent.

develop communication and collaboration skills. This difference constitutes the second contrast Watanabe noted. Lower-track classes missed out on these opportunities to learn for the sake of test preparation. The third difference was that upper-track classes had more reading and writing practice, both in class and at home. In the upper-track classes, students not only complete more readings, but also write more essays, oftentimes starting them in class and completing them at home. By contrast, the lower-track classes receive a simplified version of a smaller portion of the course content and complete worksheets, rather than write essays, to check for understanding. Fourth, and along the same lines, assignments and instruction in the upper-track classes are more challenging. Beyond differences in the reading difficulty of some of the assigned texts, the questions posed to the upper-track classes required higher-level thinking, compared to those posed to students in the lower tracks. Compared to upper-track classes, the discussions and assignments in the lower-track classes broke concepts down into smaller pieces, worked step-by-step, and emphasized the students' own experiences more than the content of the readings. Finally, Watanabe observed that upper-track classes received more written and immediate feedback on essay assignments than the lower-track classes. In one stark example, a teacher wrote comments on the essays of 53% of the upper-track students and only 13% of the lower-track students. In the other class, five randomly-chosen student essay folders in the upper-track classes received 275 corrections while five from the lower-track classes received only 58 corrections.

As Watanbe states, "These differences in instructional opportunity have far-reaching implications for students' educational and career trajectories and their future roles as citizens…[Upper-track classes equip students] with high-status knowledge and skills that prepare them to go to college, take positions of leadership in the workplace and to participate actively as informed citizens in our democracy. Meanwhile, students in the ‘regular’ track do not have equal access to this knowledge and skill set” (pp. 523). The author’s policy recommendations focus on the issue of tracking and put forth convincing arguments for the detracking of classrooms. The line of argument relies on the incentive structure provided by high-stakes testing, but she is not nearly as condemning of those policies as she is of tracking. Nonetheless, the implication is clear: while accountability policies may be credited with getting different tracks to cover the same content, there remains a substantial divide in how that content is presented, taught, and used. High-stakes accountability policies in North Carolina, Watanabe concludes, actually widens the gap in the opportunities to learn between upper- and lower-track classes, and by default widens the gap between White and Black and rich and poor. Oakes, who literally wrote the book on the negative consequences of tracking, agrees that high-stakes tests only exacerbate the achievement gaps fueled by tracking: “The policy levers of accountability tests appear, at best, largely ineffective and most likely contradictory to the policy aims” for which they are initiated (Oakes, 2008, pp. 707).

CHAPTER 7: CONCLUSION

The previous six chapters of this dissertation pursue a line of inquiry that explores achievement gaps that persist and expand in the context of high-stakes testing and academic tracking. Existing literature and the findings in this study support the theory of Effectively Maintained Inequality which states that education reforms aimed at improving educational equity are doomed to fail because privileged families will find ways to preserve their advantage and secure educations for their children that are both quantitatively and qualitatively better than the educations others receive. To perform the statistical analysis I used a unique longitudinal dataset that followed one cohort of North Carolina students from middle school through high school and college. To support my conclusions, I drew from qualitative studies in the literature that pursued a similar line of inquiry. This chapter reviews the most important elements of the previous chapters, acknowledges limitations of the study, and makes recommendations for policy and future studies.

The promise of high-stakes accountability policies was the provision of incentives for educators to push their students to achieve more. Doing so, it was (and still is) believed by some, would improve achievement across the board, but especially among populations that have been short-changed by concentrated poverty and other disadvantages. Perhaps some progress has been made in those areas – this particular study does not address that question. Its results do indicate, however, that students disadvantaged by race or by SES do not do as well at the next level of schooling as their performance on high-stakes tests would suggest. These findings give empirical support to something that critics of accountability policies have been saying again and again – that in a high-stakes testing environment, schools devote resources to preparing for tests, and neglect other important aspects of education that may be more important for the growth, development, and preparation of students, especially among the most at-risk students. If high-stakes tests become the sole objective of students, teachers, principals, and schools, only the students with privileged backgrounds will receive the broad, liberal education that prepares an individual to make the transition into and through the next level of schooling.

This project set out to answer three questions. The first asked whether North Carolina’s high-stakes test results reflect the achievement gaps that are often found in standardized tests. The answer, according to the data and analyses used in this project, is the affirmative. In both the middle school EOG tests and the high school EOC tests, students from low SES backgrounds score lower than other students. Blacks and Native Americans both score significantly below Whites and Asians, as do Hispanics in some models. All three of the theoretical approaches addressed in this dissertation state that family background plays a significant role in a child’s educational path. Where they differ is in the role background plays as students make progress through and between levels of education.

The second question asked whether these gaps changed over the transitions through middle school, high school, and college. When we hold prior academic achievement constant, we find that background characteristics of race and class reflect achievement gaps, suggesting that gaps are not merely maintained, but continue to expand. In the Life Course Perspective (LCP), as children age they become less dependent upon parents economically and socially (Müller & Karle, 1993), so background effects should wane as the child grows up. Maximally Maintained Inequality (MMI) states that background has the greatest impact at transitions into levels of education that are not universal (Raftery & Hout, 1993), which explains the gaps that appear in college GPA while controlling for high school achievement, but not the gaps in high school achievement while controlling for middle school achievement. Effectively Maintained Inequality (EMI), on the other hand, predicts that background has a powerful impact on differences in educational outcomes at all levels because even when an education level is universal, privileged families will always be able to find ways to provide for their children educational programs that are qualitatively superior to those of other children (Lucas, 2001). This dissertation's results show strong support for EMI.

The third question asked how tracking influences the impact of high-stakes tests on education. According to Lucas (1999), academic tracks are one of the key strategies by which privileged parents secure better learning experiences for their children. While my results indicate that, overall, performance on high-stakes tests is a predictor of college success, the relationship between EOC score and college GPA is only robust for those who took all of their EOC courses at the honors level. The more classes a student takes at the non-honors level, the less her EOC performance will predict success in college, even when background variables are held constant. The conclusion I draw to account for the varying degree to which EOC scores correlate with college GPA places the blame on differences in the classroom experiences made available in different academic tracks.

That conclusion is perfectly aligned with both EMI theory and recent qualitative research on teachers’ responses to high-stakes testing.

Poor Measures of Achievement

My results indicate that standardized tests are aligned with other measures of student success for some students but not others. This outcome raises the question of whether these tests are valid, an issue which is of great importance but beyond the scope of this study. I treat the tests as measures of success in school and a predictor of later success, but I do not explore what the numbers actually represent. Does a good score on an EOC test reflect the acquisition of knowledge, the mastery of facts, the development of skills, or the ability to solve problems? How about intelligence, persistence, confidence, or ambition? Does an EOC score represent hours spent studying or intellectual effort?

In the very same month that Bush was (barely) elected to the presidency of the United States, a journal article by Smith and Fey (2000) argued that accountability policies would compromise the validity of standardized tests. Part of that loss of validity comes from a variety of actors trying to game the system (Nichols & Berliner, 2007), but even before high stakes are applied, standardized tests only tell us so much.

Before tests are written, standards must be established. Generating high-quality and appropriate standards is difficult, and many states have had to settle for standards that are not terribly useful or effective (Lane, 2004). In addition, the benchmarks for what constitutes proficiency are arbitrary, in some cases because the standards themselves are arbitrary and in others because there is no clear answer to the question of how many multiple-choice questions correct out of some total number should constitute a passing grade (Neal & Schanzenbach, 2010). Although the scoring of these exams may be an objective process, the decisions about what to teach, what to test, and how many correct answers are enough to pass are all subjective decisions.

After the tests are administered and scores reported, how the results are used is another threat to validity. As mentioned above there is the possibility of intentially corrupted results, but even when everyone acts honestly there is a problem with drawing important conclusions about individuals based on standardized tests. Using results to detect areas in which a student could improve, but deciding that a particular student is proficient or not proficient based on a standardized test administered to tens of thousands of students violates the purposes for which experts say standardized tests are designed (Kornhaber, 2004). The second chapter of this dissertation addresses some of the other threats to validity in standardized and high-stakes testing and the issue deserves further research. The only assumption this study makes about the validity of EOC and EOG tests is that the degree to which they are valid varies with the context in which relevant learning takes place.

Repairing a Broken System

To some educators, students, and parents, the inequality associated with high-stakes tests and academic tracking may be perfectly acceptable. After all, higher-track courses should provide more challenge and attract more ambitious students, or else there would be no point to having tracked classes. A student who enrolls in honors biology or AP U.S. History presumably wants to cover more material and examine it more deeply than other students, and that choice should result in more academic success later on, even when performance on those tests are held constant. Accepting that notion, however, seems to contradict the standards movement, which seeks to bring consistency in curriculum and achievement to all students in all classes in all schools. How do we reconcile an interest in equal outcomes with intentionally unequal learning experiences?

Perhaps both academic tracking and high-stakes testing could be done better. In the case of the former, we need to look at the best practices that teachers use with their honors, AP, and IB students and find ways for them to bring those same sorts of activities into their lower-track classes. If analysis, synthesis, and evaluation are useful activities for learning at the honors level, they should also be effective at the non-honors level.

Skeptics of the current accountability policies that evaluate students, teachers, and schools based on the results of high-stakes tests argue that teachers of the least-advantaged students are the ones under the most pressure to narrow the curriculum, teach to the test, and "drill and kill" to the point that these students receive a substandard education designed only to reach an arbitrary level of "proficiency" (Neal & Schanzenbach, 2010). Clearly, such incentives are not helpful in efforts to close the achievement gap or to improve overall academic achievement. Considering previous research on the topic and the results of this dissertation, several improvements to policy seem warranted:

1) Bring the best practices of education to all students in all classrooms. If tracking is politically necessary, provide training and incentives that will inspire teachers to teach their lower-track classes with the sort of higher-level thinking skills usually found only in upper-level classes. In other words, give different academic tracks similar challenges, not just similar content. Analysis, synthesis, and

evaluation are not skills that are necessary only for the most elite students – every educational and occupational course benefits from them.

2) Redesign high-stakes tests so that they reflect broad skills, not narrow knowledge. When I taught at a North Carolina public school, my department distributed to all students a document called “100 facts for the Civics/Economics EOC.” The message we were sending to students was clear: for those of you who won’t bother studying more than the absolute minimum for your final exam, learn these facts and you just might pass. It is not the message you send to inspire enthusiasm for the material or confidence in abilities, let alone prepare students for the academic demands of higher education. Rather than whittling down a curriculum to the absolute minimum, what if the accountability tests instead gave students the opportunity to demonstrate their progress on their own terms, rather than in response to a grab-bag of multiple choice questions requiring mere recognition?
3) Make the tests educationally meaningful. Tests are useful tools, but the way states have developed their testing programs under NCLB and other accountability programs puts the cart before the horse. Rather than identifying strengths and weaknesses in students so that teachers, parents, and students can harness abilities and address deficiencies, our system evaluates everyone on the job they’ve done and then cuts them lose for the summer only to be assigned to new classes two months later. A more formative testing program would give everyone involved a profile of the learner so that teaching could be tailored to the student’s needs rather than to the 100 facts most likely to appear on the EOC at the end of the year.

4) Make schools and classrooms more diverse. One of the problems with accountability programs, as they are currently conceived, is that they pit schools against schools. Schools and classrooms were becoming more segregated even before high-stakes tests and the pass rates of schools became yet another method for middle class families to determine which schools and districts to avoid and which ones to choose, creating even more fuel for segregation.
5) Clarify the difference between high and low tracks and remove the GPA bump for honors classes. My experience teaching in five different high schools across three different states tells me that educators and students are unable to give a clear or consistent explanation of the difference between honors and regular classes. There is general agreement that honors classes are harder, but no firm benchmark for the honors label seems to exist. As one teacher once told me, “honors work is like pornography; I know it when I see it.” Advanced Placement and International Baccalaureate have organizations defining their standards independent of the schools, but “honors” can mean whatever the district, school, or teacher decides it will mean. The vague distinction often means that the selection of an honors class means the assigning of credentials to certain students without any real criteria. Not only do the students who enroll in honors classes get recognition for it on their transcript, but they also receive an increase in their GPA. In every school where I have ever taught, a B in an honors class is worth the same as an A in a regular class in the calculation of GPA. Removing that bonus credit from the honors class or laying out specific requirements for all honors students would help to make honors a mark of academic drive and skill rather than just a badge that

> gets handed out to the students who sign up for the classes that may or may not have substantively more challenging or even relevant demands on the students.

The policy suggestions listed above could be implemented without much cost or administrative difficulty. Realistically, the greatest obstacles would be political in nature and would come from upper middle class parents. The theory of Effectively Maintained Inequality states that parents with sufficient resources will use them to ensure that their children receive educational opportunities and outcomes superior to other children. As a result, the potential for reforms to improve equity is always in doubt. My first suggestion, to bring best practices to lower-track classes, would not be opposed by elite families, but if it were implemented, those families would quickly find ways to make their children's academic tracks somehow superior. Nonetheless, bringing those best practices to lower-track students would significantly improve their educational opportunities.

My second and third suggestions, to redesign the tests to reflect broad skills and use them as formative assessments, would draw a similar response. Privileged parents would see to it that either their children developed better skills or obtain other credentials that would give them more advantage, but again, if the outcome is an better opportunities for low-track students than currently exist under a narrow, content-based exam based on low-level cognitive skills, low-track students will be better off and will stand a better chance at the next level of education.

The fourth and fifth suggestions would be likely to be met with the most resistance, because they would dismantle two of the most concrete symbols of educational privilege, segregation and extra credit. Battles have waged for years over policies that attempt to make schools and classrooms more diverse. As my conversation with the former governor illustrates, even the most ardent supporters of integrated schools stop short of fighting to integrate classrooms. Tracking may be the political concession that has to be made to get middle class families to the table in discussions about integration policies, but then what good will diversity at the school level really achieve? And while elite parents may not be as concerned about the elimination of the GPA bump associated with upper track classes, their privileged children certainly will be. The principal or board member who proposes eliminating the weighted GPA from student transcripts will find an angry crowd of the school's brightest, wealthiest, and most popular students vehemently opposing the plan. Nonetheless, these two changes, if successfully implemented, could do a lot to improve educational equity.

Limitations

The analytic steps laid out above have some limitations. Several variables, as I have alluded to previously, could be made more precise. The SES variable, for example, is clumsy in that the measures of FRL and high school-dropout parents are likely to miss some students who come from low-SES backgrounds. The mean EOC scores used may likewise be imprecise in cases for which students took very few of the tests (and therefore their score reflects a smaller range of skills and experiences) or took them in different schools (thus possibly reflecting the effects of more than one school as well as the impact of transferring schools) or took them especially early or late in their high school years (thus being influenced by differences in maturation and school effects). Future research could explore the separate content areas of the EOC and EOG tests and the effects of different schools. The former may yield some interesting new information about the way high-stakes tests affect learning within different disciplines. The latter might be illuminating as well, but the field of education policy already has a lot of research that examines these school differences.

The design of this study relies on college performance as an outcome. Using only the UNC-bound students for those analyses risks some problems associated with selection bias as well as the exclusion of other important outcomes. While preparation for college is an extremely important task for high schools, it is not their only charge. High schools must also prepare students for work, adulthood, citizenship, and other aspects of life. This study does not assess any of those outcomes, mostly because those outcomes are so difficult to track and measure. Furthermore, a college GPA may not be the most reliable measure of academic achievement. As some current research has indicated, the level of academic performance required by many colleges is inconsistent at best, and at worst is disturbingly low (Arum & Roksa, 2010).

Another shortcoming is the lack of a control group. The data allow me to compare races, social classes, achievement, and other measures, but it does not allow me to compare courses with high-stakes tests to courses without high-stakes tests. The reason for this is obvious: the state maintains the most extensive records on the courses for which it requires standardized tests. We could develop a much better understanding of how high-stakes tests affect learning and teaching if the same data (including low-stakes test results, which don’t exist) were available for non-state-tested classes.

Similarly, there is no control group for tracking. Although some schools have successfully detracked their classes, they are too few in number and too unique to be useful in an empirical study of this size. Without a large number of public schools that vary in resources, size, performance, and diversity, it is difficult to say for certain whether tracking is the main cause of the growing achievement gaps.

I have already described some of the literature that fills in some of that gap. Oakes' time-tested *Keeping Track* (2005) gives a thorough analysis of the unequal opportunities to learn associated with tracking from even before the era of high-stakes tests, and in her more recent work she finds that accountability policies have turned education into a "scarce commodity" for which communities and families must compete. High-stakes testing, in particular, labels some students as undeserving of those scarce educational resources (Oakes, 2008). Watanabe's case study of tracking and EOG testing in a North Carolina middle school offers persuasive evidence that teachers respond to the state's high-stakes tests by increasing the number of test-related and basic skills-focused lessons in their regular classes while developing more higher-order thinking skills in their honors classes. That study and its findings are described in Chapter 6.

Future Research

Future research could help develop the above suggestions more thoroughly and address some of the limitations. How can students with low motivation or deficiencies in skills be guided through challenging course material and assignments? How can a test maintain validity and reliability while also permitting students of diverse family and educational backgrounds sufficient flexibility to demonstrate their mastery? What could teachers and students do with the results of a formative assessment given at the start of the school year? What are effective policies for assigning the label "honors" to a class or a student that would reflect achievement more than background and stop short of distorting academic records by favoring those who are already advantaged? How can diversity in the classroom be made more palatable to families who feel they lose something valuable when their child learns alongside a child of a different class, race, or ability level?

These inquiries all hint at a larger question about what a more equitable public education system would look like. Even if institutional realities require that we merely “tinker toward utopia”, having a good idea of what the final product should look like would be beneficial. This dissertation does not include that prescription, but it does offer some ideas of what not to do. A school system with true equity would not funnel some students into paths rich with learning opportunities and others onto paths devoid of inspiration or lacking in the development of skills, abilities, and confidence. It would not use competition as the solution to gaps in achievement associated with class or racial minority status. Instead, it would provide every school and student with high standards, ample chances to demonstrate unique competence and growth, and resources necessary to provide abundant opportunities to learn.

Contributions of this Study

This dissertation contributes to the literature on education policy in four ways. First, it provides measurements of achievement gaps associated with class and race. While not unexpected, the findings are important because they show that a decade after high-stakes tests were put into effect in North Carolina, family background continues to be a strong predictor of student outcomes. As policy makers continue to wrestle with ideas about how to improve student achievement, an effective solution will be one that recognizes the significant role that social class and race continue to play in determining student success.

Second, it reveals that gaps expand over the course of middle school, high school, and college. Rather than shrinking the gaps that have roots in student background, schools seem to be exacerbating them. My dissertation shows that in middle school, standardized tests show a divide associated with class and race. At high school, several of those divides have grown larger and in college they grow larger still. Rather than overcoming background effects, schools seem to be exacerbating them.

Third, it connects tracking to high-stakes testing to assess the impact of the combination of these two policies on student outcomes. It is in this area that my dissertation makes its greatest contribution. We know from other studies that under accountability policies utilizing high-stakes tests, educators narrow the curriculum as well as the range of classroom experiences. We know from research on tracking that lower-track students receive fewer and lower-quality opportunities to learn. Watanabe (2008) observed the intersection of high-stakes tests and academic tracking and reported that the negative consequences of high-stakes tests are especially pronounced in lower-track classes. In her study, upper-track students received far less test-preparation instruction and far more higher-order thinking activities. My study uses large-scale quantitative analysis to trace the impact of the mechanism Watanabe observed. Unfortunately for the cause of educational equity, I find that the shift from other activities to test preparation that is likely to be found in lower-track classes may be successful at raising those students’ test scores, but it does so at the cost of insufficient readiness for the next level of education. My dissertation, therefore, provides quantifiable evidence that high-stakes testing is making educational equity more difficult to achieve and doing so, at least in part, through the existing structure of academic tracks within schools.

Such findings are important for policy makers because they expose not only the limitations of high-stakes as an accountability strategy but also academic tracking as another source if inequity. Neither one is a new finding on its own, but bringing the two issues together is important for seeing the complexity of the issue of school reform. If we only consider one aspect of a reform and ignore the manner in which it plays out with other characteristics of public education, we only ever "tinker" around the edges and never establish the education system our children need and deserve.

Finally, this dissertation gives strong support to Lucas's theory of Effectively Maintained Inequality (EMI). More than a decade ago Lucas argued that parents with sufficient resources were using academic tracking a means to procure for their children a public education that was in many ways better than the educational opportunities afforded to other children. My findings reveal why middle- and upper-class families fight to institute and preserve academic tracking policies in their children's schools. More than just an effort to segregate students within diverse schools, the practice of tracking results in substantial differences in educational outcomes.

High-stakes tests will not bring educational inequality to an end. Creating an environment in which students must either sink or swim does little to change the fact that some students are carrying more weight and others are getting extra lessons. Standardized tests do not level the playing field as much as they continue the ongoing separation of students into categories of success and failure. Academic tracking sets the stage for that separating to take place. For those who are fortunate enough to progress through middle and high school in upper-track classes, lessons will provide an easy balance between preparation for the tests and preparation for the future. Those that earn high grades in their honors high school classes will do so by doing well on the test prep activities as well as the higher-level thinking strategies that the teachers of such classes typically provide. In lower-track classes, students and teachers will work very hard to make sure everyone hits the mark to earn the proficiency mark, but that goal will distract them from the other skills and content they should be learning. When each arrives in college, the students who took honors classes will be prepared, while their peers who thought they would play it safe in their high school course selection, lacked confidence in their abilities, were discouraged by someone from stretching themselves, or simply could not fit honors courses into their schedules will find college more difficult – even if they score just as well on the state's high-stakes tests.

REFERENCES

Amrein, A., & Berliner, D. (2002). High-Stakes Testing, Uncertainty, and Student Learning. Education Policy Analysis Archives, 10(8). Retrieved from http://epaa.asu.edu/epaa/v10n18/

Apple, M. W. (1995). Education and Power. Psychology Press.

Arum, R., & Roksa, J. (2010). Academically Adrift: Limited Learning on College Campuses. University of Chicago Press.

Atkinson, R. C. (2002). Achievement versus Aptitude in College Admissions. Issues in Science and Technology, 18(2), 31–36.

Attewell, P. (2001). The Winner-Take-All High School: Organizational Adaptations to Educational Stratification. Sociology of Education, 74(4), 267. doi:10.2307/2673136

Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. Educational Researcher, 36(5), 258–267. doi:10.3102/0013189X07306523

Au, W. (2009). Unequal By Design: High-Stakes Testing and the Standardization of Inequality. Taylor & Francis.

Berliner, D. (2011). Rational responses to high stakes testing: the case of curriculum narrowing and the harm that follows. Cambridge Journal of Education, 41(3), 287–302. doi:10.1080/0305764X.2011.607151

Birk, L. (2001). Collateral Damage. Harvard Education Letter, 17(2).

Boaler, J., & Staples, M. (2008). Creating Mathematical Futures through an Equitable Teaching Approach: The Case of Railside School. Teachers College Record, 110(3), 608–645.

Booher-Jennings, J. (2005). Below the Bubble: “Educational Triage” and the Texas Accountability System. American Educational Research Journal, 42(2), 231–268. doi:10.3102/00028312042002231

Bowles, S., & Gintis, H. (1976). Schooling in Capitalist America: Educational Reform and the Contradictions of Economic Life. Routledge and Kegan Paul.

Bowles, S., & Gintis, H. (2002). The Inheritance of Inequality. Journal of Economic Perspectives, 16(3), 3–30. doi:10.1257/089533002760278686

Braun, H., Chapman, H., & Vezzu, S. (2010). The Black-White Achievement Gap Revisited. Education Policy Analysis Archives, 18(21). Retrieved from http://epaa.asu.edu/ojs/article/view/772.

Brennan, R. T., Kim, J., Wenz-Gross, M., & Siperstein, G. N. (2001). The Relative Equitability of High-Stakes Testing versus Teacher-Assigned Grades: An Analysis of the Massachusetts Comprehensive Assessment System (MCAS). Harvard Educational Review, 71(2), 173–216.

Brown, A. B., & Clift, J. (2010). No Child Left Behind Act of 2001. Product Page. Retrieved June 5, 2012, from http://www.rand.org/pubs/external_publications/EP20100059.html

Burris, C. C., Wiley, E., Welner, K. G., & Murphy, J. (2008). Accountability, Rigor, and Detracking: Achievement Effects of Embracing a Challenging Curriculum as a Universal Good for All Students. Teachers College Record, 110(3), 571–607.

Bush, G. W. (2001). No Child Left Behind. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED447608

Campbell, D. T. (1979). Assessing the impact of planned social change. Evaluation and Program Planning, 2(1), 67–90. doi:10.1016/0149-7189(79)90048-X

Carnoy, M. (2003). The New Accountability: High Schools and High-Stakes Testing. Psychology Press.

Clotfelter, C. T. (2010). American Universities in a Global Market. University of Chicago Press.

Cohn, E., Cohn, S., Balch, D. C., & Bradley Jr., J. (2004). Determinants of undergraduate GPAs: SAT scores, high-school GPA and high-school rank. Economics of Education Review, 23(6), 577–586. doi:10.1016/j.econedurev.2004.01.001

Corbett, C., Hill, C., & St. Rose, A. (2008). Where the Girls Are: The Facts about Gender Equity in Education. Executive Summary. American Association of University Women Educational Foundation. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED501320

Costrell, R., & Peyser, J. (2004). Exploring the Costs of Accountability. Education Next, 4(2), 22–29.

Darling-Hammond, L. (2004). The Color Line in American Education: Race, Resources, and Student Achievement. Du Bois Review: Social Science Research on Race, 1(02), 213–246. doi:10.1017/S1742058X0404202X

Darling-Hammond, L. (2010). The flat world and education: how America’s commitment to equity will determine our future. Teachers College Press.

Darling-Hammond, L., Ancess, J., & Falk, B., (1995). Authentic Assessment in Action: Studies of Schools and Students at Work. Teachers College Press.

Diamond, J. B., & Spillane, J. P. (2004). High-Stakes Accountability in Urban Elementary Schools: Challenging or Reproducing Inequality? Teachers College Record, 106(6), 1145–1176.

Dillon, S. (2006, March 26). Schools Cut Back Subjects to Push Reading and Math. The New York Times. Retrieved from http://www.nytimes.com/2006/03/26/education/26child.html

Farkas, G. (2003). Racial Disparities and Discrimination in Education: What Do We know, How Do We Know It, and What Do We Need to Know? Teachers College Record, 105(6), 1119–1146. doi:10.1111/1467-9620.00279

Ferguson, R. F. (1998). Teachers' Perceptions and Expectations and the Black-White Test Score Gap. The Black-White Test Score Gap. Brookings Institution Press.

Gamoran, A. (2001). American Schooling and Educational Inequality: A Forecast for the 21st Century. Sociology of Education. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=EJ679984

Gamoran, A. (2009). Tracking and Inequality: New Directions for Research and Practice. WCER Working Paper No. 2009-6. Wisconsin Center for Education Research. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED506617

Gamoran, A., Nystrand, M., Berends, M., & LePore, P. C. (1995). An Organizational Analysis of the Effects of Ability Grouping. American Educational Research Journal, 32(4), 687–715. doi:10.3102/00028312032004687

Geiser, S. (2009). Back to the Basics: In Defense of Achievement (and Achievement Tests) in College Admissions. Change: The Magazine of Higher Learning, 41(1), 16–23.

Geiser, S., & Santelices, M. V. (2007). Validity of High-School Grades in Predicting Student Success beyond the Freshman Year: High-School Record vs. Standardized Tests as Indicators of Four-Year College Outcomes. Research & Occasional Paper Series: CSHE.6.07. Center for Studies in Higher Education. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED502858

Geiser, S., & Studley, with R. (2002). UC - and the SAT: Predictive Validity and Differential Impact of the SAT - I and SAT - II - at the University of California. Educational Assessment, 8(1), 1. doi:10.1207/S15326977EA0801_01

Gibson, M. A. (1988). Accommodation Without Assimilation: Sikh Immigrants in an American High School. Cornell University Press.

Gordon, R., Piana, L. D., Burlingame, P., Center, A. R., Monifa, A., & Wang, O. (1999). No Exit?: Testing, Tracking, and Students of Color in U.S. Public Schools. Applied Research Center.

Greene, J. P., Winters, M. A., & Forster, G. (2003). Testing High Stakes Tests: Can We Believe the Results of Accountability Tests? Civic Report. Manhattan Institute for Policy Research. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED475488

Grissmer, D., & Flanagan, A. (1998). Exploring Rapid Achievement Gains in North Carolina and Texas. Lessons from the States. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED425204

Grundy, S. (1987). Curriculum: Product Or Praxis? Falmer Press.

Gunn, H. E., & Singh, J. (2004). Minority Report: How African Americans and Hispanics Can Increase Their Test Scores. R&L Education.

Haney, W. (2000). The Myth of the Texas Miracle in Education. Education Policy Analysis Archives, 8, 41.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2009). New Evidence about Brown v. Board of Education: The Complex Effects of School Racial Composition on Achievement. Journal of Labor Economics, 27(3).

Hanushek, E. A., & Woessmann, L. (2010). The Economics of International Differences in Educational Achievement (Working Paper No. 15949). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w15949

Harris, D. M., & Anderson, C. R. (2012). Equity, Mathematics Reform and Policy: The Dilemma of “Opportunity to Learn.” In B. Herbel-Eisenmann, J. Choppin, D. Wagner, & D. Pimm (Eds.), Equity in Discourse for Mathematics Education, Mathematics Education Library (Vol. 55, pp. 195–204). Springer Netherlands. Retrieved from http://www.springerlink.com/content/u251512038250260/abstract/

Harris, D. N., & Herrington, C. D. (2006). Accountability, Standards, and the Growing Achievement Gap: Lessons from the Past Half-Century. American Journal of Education, 112(2), 209–238. doi:10.1086/498995

Harwell, M., & LeBeau, B. (2010). Student Eligibility for a Free Lunch as an SES Measure in Education Research. Educational Researcher, 39(2), 120–131. doi:10.3102/0013189X10362578

Helms, A. D. (2012, August 25). CMS primed for new standards in new year. Charlotte Observer. Retrieved from http://www.charlotteobserver.com/2012/08/25/3478345/cms-primed-for-new-standards.html

Henig, J. R., Hula, R. C., Orr, M., & Pedescleaux, D. S. (2001). The Color of School Reform: Race, Politics, and the Challenge of Urban Education. Princeton University Press.

Hess, F. M. (1999). Spinning Wheels: The Politics of Urban School Reform. Brookings Institution Press.

Hess, F. M. (2006). Common Sense School Reform. Macmillan.

Heubert, J. P., & Hauser, R. M. (1999). High Stakes: Testing for Tracking, Promotion, and Graduation. National Academies Press.

Hoffman, J. V., Assaf, L. C., & Paris, S. G. (2001). High-Stakes Testing in Reading: Today in Texas, Tomorrow? Reading Teacher, 54(5), 482–92.

Hovey, K. A., & Hovey, H. A. (2007). CQ's State Fact Finder: Rankings Across America. CQ Press.

Howard, T. C. (2001). Telling Their Side of the Story: African-American Students' Perceptions of Culturally Relevant Teaching. The Urban Review, 33(2), 131–149. doi:10.1023/A:1010393224120

James, R. R. (2009). How to Mend a Broken Act: Recapturing Those Left Behind by No Child Left Behind. Retrieved from http://works.bepress.com/regina_james/1

Jencks, C., & Phillips, M. (1998). The Black-White Test Score Gap. Brookings Institution Press.

Johnson, A. W. (2009). Objectifying Measures: The Dominance of High-Stakes Testing and the Politics of Schooling. Temple University Press.

Jones, M. G., Jones, B. D., Hardin, B., Chapman, L., Yarbrough, T., & Davis, M. (1999). The Impact of High-Stakes Testing on Teachers and Students in North Carolina. Phi Delta Kappan, 81(3), 199–203.

Jones, M. G., Jones, B. D., & Hargrove, T. Y. (2003). The Unintended Consequences of High-Stakes Testing. Rowman & Littlefield.

Kafer, K. (2004). No Child Left Behind: Where Do We Go from Here? Backgrounder No. 1775. The Heritage Foundation. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED483841

Kane, T. J., Staiger, D. O., & Riegg, S. K. (2005). School Quality, Neighborhoods and Housing Prices: The Impacts of school Desegregation. National Bureau of Economic Research Working Paper Series, No. 11347. Retrieved from http://www.nber.org/papers/w11347

King, G., Keohane, R. O., & Verba, S. (1994). Designing Social Inquiry: Scientific Inference in Qualitative Research. Princeton University Press.

Kirst, M. W., & Venezia, A. (2004). From High School to College: Improving Opportunities for Success in Postsecondary Education. Jossey-Bass, An Imprint of Wiley. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED496375

Klugman, J. (2011). The Advanced Placement Arms Race and the Reproduction of Educational Inequality. Retrieved from http://citation.allacademic.com/meta/p_mla_apa_research_citation/5/0/8/0/6/p508060_index.html

Klugman, J. (2012). How Resource Inequalities Among High Schools Reproduce Class Advantages in College Destinations. Research in Higher Education. doi:10.1007/s11162-012-9261-8

Kornhaber, Mindy L. (2004). Appropriate and Inappropriate Forms of Testing, Assessment, and Accountability. Educational Policy, 18(1), 45–70. doi:10.1177/0895904803260024

Kornhaber, Mindy Laura. (1997). Seeking Strengths: Equitable Identification for Gifted Education and the Theory of Multiple Intelligences. Harvard Graduate School of Education.

Kozol, J. (1992). Savage inequalities: children in America’s schools. HarperCollins.

Ladd, H. F., & Zelli, A. (2002). School-Based Accountability in North Carolina: The Responses of School Principals. Educational Administration Quarterly, 38(4), 494–529. doi:10.1177/001316102237670

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. Biometrics, 38(4), 963–974.

Lane, S. (2004). Validity of High-Stakes Assessment: Are Students Engaged in Complex Thinking? Educational Measurement: Issues and Practice, 23(3), 6–14. doi:10.1111/j.1745-3992.2004.tb00160.x

Lareau, A. (1987). Social Class Differences in Family-School Relationships: The Importance of Cultural Capital. Sociology of Education, 60(2), 73. doi:10.2307/2112583

Lareau, A. (2003). Unequal childhoods: class, race, and family life. University of California Press.

Lee, J., & Wong, K. K. (2004). The Impact of Accountability on Racial and Socioeconomic Equity: Considering Both School Resources and Achievement Outcomes. American Educational Research Journal, 41(4), 797–832. doi:10.3102/00028312041004797

Lee, V. E., & Burkam, D. T. (2002). Inequality at the Starting Gate: Social Background Differences in Achievement as Children Begin School. Economic Policy Institute. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED470551

Leonardo, Z. (2007). The war on schools: NCLB, nation creation and the educational construction of whiteness. Race Ethnicity and Education, 10(3), 261–278. doi:10.1080/13613320701503249

Levin, H. (2001). Waiting for Godot: Cost-Effectiveness Analysis in Education. New Directions for Evaluation, 2001(90), 55–68. doi:10.1002/ev.12

Linn, R. L. (2000). Assessments and Accountability. Educational Researcher, 29(2), 4–16. doi:10.3102/0013189X029002004

Logan, J. R., Oakley, D., & Stowell, J. (2008). School Segregation in Metropolitan Regions, 1970–2000: The Impacts of Policy Choices on Public Education. American Journal of Sociology, 113(6), 1611–1644. doi:10.1086/587150

Logan, J. R., Stowell, J., & Oakley, D. (2002). Choosing Segregation: Racial Imbalance in American Public Schools, 1990-2000. Lewis Mumford Center for Comparative Urban and Regional Research. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED471516

Lucas, Samuel R. (2001). Effectively Maintained Inequality: Education Transitions, Track Mobility, and Social Background Effects. The American Journal of Sociology, 106(6), 1642–1690.

Lucas, Samuel Roundfield. (1999). Tracking Inequality: Stratification and Mobility in American High Schools. Sociology of Education Series. Teachers College Press. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED447225

Madaus, G. F., & Clarke, M. (2001). The Adverse Impact of High Stakes Testing on Minority Students: Evidence from 100 Years of Test Data. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED450183

McDermott, K. A. (2007). “Expanding the Moral Community” or “Blaming the Victim”? The Politics of State Education Accountability Policy. American Educational Research Journal, 44(1), 77–111. doi:10.3102/0002831206299010

McNeil, L. (2005). Faking Equity: High-Stakes Testing and the Education of Latino Youth. In A Valenzuela (Ed.), Leaving Children Behind: How “Texas-Style” Accountability Fails Latino Youth (pp. 57–111). New York: State University of New York Press.

McNeil, L. M. (2000). Creating New Inequalities: Contradictions of Reform. Phi Delta Kappan, 81(10), 728–34.

Medina, N., & Neill, D. M. (1988). Fallout from the Testing Explosion: How 100 Million Standardized Exams Undermine Equity and Excellence in America’s Public Schools. FairTest. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED301580

Mickelson, R. A. (2001). Subverting Swann: First- and Second-Generation Segregation in the Charlotte-Mecklenburg Schools. American Educational Research Journal, 38(2), 215 –252. doi:10.3102/00028312038002215

Mickelson, R. A., & Everett, B. J. (2008). Neotracking in North Carolina: How High School Courses of Study Reproduce Race and Class-Based Stratification. Teachers College Record, 110(3), 535–570.

Milner, M. (2004). Freaks, Geeks, and Cool Kids: American Teenagers, Schools, and the Culture of Consumption. Psychology Press.

Moe, T. M. (2002). Politics, Control, and the Future of School Accountability. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED477176

Müller, W., & Karle, W. (1993). Social Selection in Educational Systems in Europe. European Sociological Review, 9(1), 1–23.

Neal, D., & Schanzenbach, D. W. (2010). Left Behind by Design: Proficiency Counts and Test-Based Accountability. Review of Economics and Statistics, 92(2), 263–283. doi:10.1162/rest.2010.12318

Nichols, S. L., & Berliner, D. C. (2007). Collateral damage: how high-stakes testing corrupts America’s schools. Harvard Education Press.

NC Department of Public Instruction (2012). K-12 Curriculum and Instruction/Standard Course of Study. Available at http://www.dpi.state.nc.us/curriculum/

Oakes, J. (1986). Keeping Track: How Schools Structure Inequality. Yale University Press.

Oakes, J. (2005). Keeping Track: How Schools Structure Inequality. Yale University Press.

Oakes, J. (2008). Keeping Track: Structuring Equality and Inequality in an Era of Accountability. Teachers College Record, 110(3), 700–712.

Olson, L. (2004). NCLB Law Bestows Bounty on Test Industry. Education Week, 24(14), 19.

Orfield, G., & Wald, J. (2000). Testing, Testing. Nation, 270(22), 38–40.

Orfield, Gary, & Lee, C. (2005). Why Segregation Matters: Poverty and Educational Inequality. Harvard Education Publishing Group. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED489186

Orr, A. J. (2011). Gendered Capital: Childhood Socialization and the "Boy Crisis" in Education. Sex Roles, 65(3-4), 271–284. doi:10.1007/s11199-011-0016-3

Packard, W. K. (1997). Sound, Basic Education: North Carolina Adopts an Adequacy Standard in Leandro v. State. North Carolina Law Review, 76, 1481.

Page, R. (1989). The Lower-Track Curriculum at a "Heavenly" High School: "Cycles of Prejudice." Journal of Curriculum Studies, 21(3), 197–221.

Page, R. N. (1991). Lower-Track Classrooms: A Curricular and Cultural Perspective. Teachers College Press. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED355324

Pearl, A. (2002). The Big Picture: Systemic and Institutional Factors in Chicago School Failure and Success. In R. Valencia (Ed.), Chicano School Failure and Success: Past, Present, and Future (2nd ed., pp. 335–364). London: Routledge Press.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, Design, and Analysis: An Integrated Approach. Psychology Press.

Perna, L. W., & Thomas, S. L. (2009). Barriers to College Opportunity The Unintended Consequences of State-Mandated Testing. Educational Policy, 23(3), 451–479. doi:10.1177/0895904807312470

Rabe-Hesketh, S., & Skrondal, A. (2012). Multilevel and Longitudinal Modeling Using Stata, 3rd Edition (Stata Press books). StataCorp LP. Retrieved from http://econpapers.repec.org/bookchap/tsjspbook/mimus2.htm

Raftery, A. E., & Hout, M. (1993). Maximally maintained inequality: expansion, reform, and opportunity in Irish education, 1921-75. Sociology of Education, 66(1), 41–62.

Ravitch, D. (2001). Left Back: A Century of Battles Over School Reform. Simon and Schuster.

Ravitch, D. (2010). The Death and Life of the Great American School System: How Testing and Choice Are Undermining Education. Basic Books.

Reville, S. P. (2004). High Standards + High Stakes = High Achievement in Massachusetts. Phi Delta Kappan, 85(8), 591.

Roderick, M., & Engel, M. (2001). The Grasshopper and the Ant: Motivational Responses of Low-Achieving Students to High-Stakes Testing. Educational Evaluation and Policy Analysis, 23(3), 197–227. doi:10.3102/01623737023003197

Rothstein, J. M. J. M. (2004). College performance predictions and the SAT. Journal of Econometrics, 121(1-2), 297–317. doi:10.1016/j.jeconom.2003.10.003

Rothstein, R., Jacobsen, R., & Wilder, T. (2008). Grading Education: Getting Accountability Right. Economic Policy Institute.

Sacks, P. (2001). Standardized Minds: The High Price of America’s Testing Culture And What We Can Do To Change It. Da Capo Press.

Singer, J. D. (1998). Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. Journal of Educational and Behavioral Statistics, 23(4), 323–355. doi:10.3102/10769986023004323

Skrla, L., & Scheurich, J. J. (2004). Educational Equity and Accountability: Paradigms, Policies and Politics. Psychology Press.

Smith, M. L., & Fey, P. (2000). Validity and Accountability in High-Stakes Testing. Journal of Teacher Education, 51(5), 334–344. doi:10.1177/0022487100051005002

Tobin, K., Roth, W.-M., & Zimmermann, A. (2001). Learning To Teach Science in Urban Schools. Journal of Research in Science Teaching, 38(8), 941–64.

Townsend, B. L. (2002). “Testing While Black” Standards-Based School Reform and African American Learners. Remedial and Special Education, 23(4), 222–230. doi:10.1177/07419325020230040501

Tyack, D. B., & Cuban, L. (1995). Tinkering toward utopia: a century of public school reform. Harvard University Press.

Valencia, R. R. (1997). The Evolution of Deficit Thinking: Educational Thought and Practice. The Stanford Series on Education and Public Policy. Falmer Press, Taylor & Francis Inc. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED413139

Valenzuela, Angela. (1999). Subtractive Schooling: U.S.-Mexican Youth and the Politics of Caring. SUNY Press.

Valenzuela, Angela. (2005). Leaving Children Behind: How “Texas-Style” Accountability Fails Latino Youth. SUNY Press.

Watanabe, M. (2008). Tracking in the Era of High Stakes State Accountability Reform: Case Studies of Classroom Instruction in North Carolina. Teachers College Record, 110(3), 489–534.

Wells, A. S., & Serna, I. (1996). The politics of culture : understanding local political resistance to detracking in racially mixed schools. Harvard educational review, 66(1), 93–118.

Welner, K. G. (2001). Legal Rights, Local Wrongs: When Community Control Collides With Educational Equity. SUNY Press.

Wiliam, D. (2010). Standardized Testing and School Accountability. Educational Psychologist, 45(2), 107. doi:10.1080/00461521003703060

Yonezawa, S., Wells, A. S., & Serna, I. (2002). Choosing Tracks:“Freedom of Choice” in Detracking Schools. American Educational Research Journal, 39(1), 37–67. doi:10.3102/00028312039001037

Zwick, R., & Himelfarb, I. (2011). The Effect of High School Socioeconomic Status on the Predictive Validity of SAT Scores and High School Grade-Point Average. Journal of Educational Measurement, 48(2), 101–121. doi:10.1111/j.1745-3984.2011.00136.x