UNPACKING TRACKING: THE ROLE OF INSTRUCTION, TEACHER BELIEFS AND SUPPLEMENTAL COURSES IN THE RELATIONSHIP BETWEEN TRACKING AND STUDENT ACHIEVEMENT

By

Rebecca Anne Schmidt

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Leadership and Policy Studies

May, 2013

Nashville, Tennessee

Approved:

Professor Thomas M. Smith

Professor Kara Jackson

Professor Christopher Loss

Professor Ronald Zimmer

DEDICATION

To Andy and Jenny. Look what you did.

ACKNOWLEDGEMENTS

This work would not have been possible without the financial support of the Vanderbilt Experimental Educational Research Training (ExpERT) Pre-Doctoral Fellowship (David S. Cordray, Director; IES Grant Number R305B040110), the Harold Sterling Vanderbilt Grant and the National Science Foundation Grant that funded the Middle School Mathematics in the Institutional Setting of Teaching Project (Paul Cobb and Thomas Smith, Co-PIs; Award Nos. ESI 0554535 and 0830029). The opinions expressed are those of the author and do not represent the views of the U. S. Department of Education or the National Science Foundation. I am also deeply indebted to the members of my committee, who provided thoughtful and invaluable feedback on every stage of this process, making the final product rigorous, relevant and readable. I am particularly grateful for the guidance of my advisor, Dr. Thomas Smith, who pushed me to think deeper, never doubting my ability to get it done and get it done well.

Finally, of course, I am unbelievably lucky in my friends and family, who listened to me grumble, ignored my threats to quit and bought me drinks at every minor success. My parents gave me life and an education, and then taught me to take neither for granted; my sister's blind faith in my genius has humbled and confused me all my life. Lisa P. made me who I am somehow still shapes me from thousands of miles away. Lisa H. was the first person who said I could do it, and I actually believed her, because telling me to do it meant moving far away. But most particularly I am grateful for Jenny and Andy, whose participation in marathon "write-ins" should honestly get them co-author credits, but I'm not that generous. Andy, in case I never get another chance, {secret code}.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Most schools and districts face the problem of how to help low-achieving students and efficiently target resources, particularly given budget cuts and accountability under No Child Left Behind. One policy that has been employed is grouping students into classrooms by their measured or perceived ability—a process known as tracking. However, research has shown that course-level grouping disproportionately assigns minority and low-income students to low-track classes (e.g., Gamoran, 2009; Oakes, 2005) and may increase inequality between high and low-achieving students (e.g., Esposito, 1973; Gamoran, 1987; Oakes, 2005), without increasing overall student achievement in the school or district (e.g., Esposito, 1973; Gamoran, 2009; Kulik & Kulik, 1982; Slavin, 1990). Opponents of tracking argue that untracked classrooms, in which students are grouped heterogeneously in terms of prior achievement, are a more equitable approach. Proponents of tracking maintain that it can help teachers tailor their instruction, and both sides have argued that their approach can work with the requisite supports.

The analyses included here attempt to dig deeper into the relationship between tracking and achievement. First, although both proponents and opponents of tracking have argued that instruction may be the mechanism by which track level affects achievement, this has not been studied quantitatively in middle school mathematics. Chapter II uses a measure of high-quality instruction focused on reasoning and

justification over recall, and examines whether instructional quality varies by track level. Then, using the same measure, it tests whether this variation can explain the gap in achievement between high- and regular-track students in middle school mathematics.

Second, some opponents of tracking have argued that teachers' own views of student ability may act as a barrier to detracking efforts. These authors have argued for a developmental conception of ability, in which "ability" is not a fixed and uni-dimensional state, but rather something that develops over time and can be influenced by instruction. They have argued that this conception of ability is necessary to the success of tracking efforts, but this has not been measured in prior research. Chapter III uses a quantitative measure of teachers' views of student ability to test whether a developmental view is associated with higher student achievement, and whether these views can support the success of untracked education.

Finally, both proponents and opponents of tracking have argued for additional supports for struggling students, but there has been little research on these supports. One such support that has been growing in popularity is "double dose" instruction, in which low-achieving students receive a full additional period of mathematics. While these policies have been adopted in more schools in recent years, there are few studies on their effectiveness, and these studies have not accounted for the ways that the characteristics of the policy vary. Chapter IV examines double dose policies in thirty schools across four urban districts, first examining how a policy under the same name can look radically different across schools. Then, I examine the overall relationship between double dose and student achievement, before finally differentiating effects by the characteristics of the policy.

Each of these chapters addresses an aspect of the relationship between ability grouping and student achievement that has been understudied. As such, I attempt to "Unpack Tracking" and provide entry points for policymakers to support low-achieving students in middle school mathematics.

CHAPTER II

THE ROLE OF INSTRUCTION AS A MEDIATOR BETWEEN TRACK LEVEL AND STUDENT ACHIEVEMENT

Despite the large number of reports examining outcomes in tracked and untracked classrooms (e.g., Esposito, 1973; Gamoran, 2009; Oakes, 2005), few large-scale studies have undertaken to look at the mechanism by which tracking may help or harm students, namely the quality of teaching. Opponents of tracking have argued that students who are placed in higher tracks have more qualified teachers and a more challenging classroom environment, which exacerbates existing achievement gaps. Tracking proponents argue that separating students by ability allows teachers to more effectively target instruction to the diverse needs of students in their schools (Gamoran, 2009; Loveless 1999). In both cases, instruction is the linchpin in making tracking or de-tracking work for students, so understanding the importance of teaching is paramount.

The vast majority of the studies on the relationship between tracking and instructional quality in mathematics are small case studies using two or three teachers or schools (e.g., Boaler, 2006; Horn, 2006; McDermott, Rothenberg & Martin, 1995; Rubin, 2008; Watanabe, 2008). A few large-scale and/or quantitative studies have been carried out on the relationship between tracking and instructional practices in reading or English Language Arts classrooms. Those studies have found that instructional approaches linked to student achievement such as coherent discussion, revising activities and homework were more common in high-track than in low-track classrooms (e.g., Applebee, Langer, Nystrand & Gamoran, 2003; Gamoran, Nystrand, Berends & LePore,

1995). Studies examining mathematics instruction have generally relied on teacher self-report (Epstein and MacIver, 1992), focusing either on non-subject-specific teacher behaviors, such as classroom management, climate and teacher enthusiasm, or on curricular materials over instruction (Evertson, 1982; Oakes, 2005). The largest study of mathematics-specific instructional differences is by Villegas (1991), who examined 89 middle grades classes in six urban districts. This study looked at types of "teacher talk" used and found significant differences between high and low-track classrooms, wherein high-track classrooms had more student-initiated talk and more "academic-" and less "directive-" type talk than low-track classrooms. This study did not examine untracked settings or other indicators of high quality mathematics instruction besides talk, or account for possible pre-existing differences between students, despite its large sample.

In this chapter I first discuss the history of tracking and ability grouping in the United States as well as the findings on the relationship between tracking and student achievement, to set the stage for why instruction may be important for understanding this relationship. Next, I discuss the existing literature on the role of instruction as a mechanism connecting tracking policies to student outcomes, demonstrating that high-quality quantitative research in the area of mathematics is severely lacking. Then I discuss what constitutes "high-quality" instruction in today's mathematics classrooms. I argue that a clear definition of instructional quality is a necessary precursor to studying it as a mediator between tracking and student outcomes, and that many prior studies have failed to establish such a definition. I also review the literature on a particular conceptualization of high-quality mathematics instruction: that reflected in the National Council of Teachers of Mathematics (NCTM). Using this definition and the direction of

prior research, I address five research questions using a large multi-state dataset focused on middle-grades mathematics instruction: 1) Are there measurable differences in student achievement by track level? 2) Are there measurable differences in instructional quality between teachers in tracked and untracked settings? 3) Between high- and low-track classrooms? 4) Is rigorous mathematics instruction associated with higher student achievement? 5) Do differences in instructional quality mediate the relationship between track level (high versus regular track) and student achievement?

Literature Review

History of Tracking. "Tracking" is a word that has been used to describe a wide variety of policies and behaviors in schools. In this analysis it refers to sorting students at a classroom level by measured or perceived ability. The methods used to assess students' "ability" have varied widely across time and between schools (Oakes, 2005). These methods include IQ and other standardized test scores; teacher, counselor and parent recommendations (often based on behavior and perceived effort); and prior course grades (Oakes, 2005). Using these measures, students can be assigned to different school types, different courses of study, different levels of individual course subjects, or different subgroups within the same classroom. Dupriez, Dumay and Vause (2008) classify countries by the selectivity of their public school systems: more selective systems separate students into ability groups at a younger age (e.g., age 10 in England), while more "comprehensive" systems may teach a common curriculum until age 16 or later.

As Resnick and Resnick (1985) outlined in their history of this issue, the level of selectivity and the types of ability grouping used in the United States have varied over time and have been driven both by historical changes and by the impact of research

findings. Before the twentieth century, the vast majority of Americans did not attend secondary school. Between 1880 and 1930 the proportion of 14 to 17-year-olds attending school in the U.S. increased dramatically: from 10% to 70% in 50 years. This was influenced in part by changes in child labor and compulsory schooling laws, both of which reflected changing social norms about the roles and responsibilities of children. This period was also marked by an ongoing tension between a national curriculum and different programs of study. In 1893, the Committee of Ten, chaired by Harvard president Charles Eliot, launched the first debate in the US over curriculum. The Committee focused on the fact that there was no consensus over what should be taught, when, or how in secondary schools, and provided the radical suggestion of a common core curriculum: science, math, languages, English and history taught for several years of increasing difficulty and meeting several times a week. (Resnick & Resnick, 1985).

During this same time there was also a growth in vocational education in response to the expansion of jobs requiring clerical and vocational skills. Vocational education grew from a few electives offering agricultural education to farmers to a "full blown vocational education movement" (Resnick & Resnick, 1985: p. 7). The 1917 Smith-Hughes act provided funding for a half-day shop program as part of vocational education that resulted necessarily in tracking: students participating in this program did not have the time in their schedules to participate in college-preparatory core courses (Resnick & Resnick, 1985). In 1918, the Cardinal Principles of Secondary Education were put out by the National Education Association, which supported differentiation of secondary education by intended vocation. These principles argued that the time to decide on

vocation was in the ages from 12 to 14, and thus signaled the birth of junior high and middle schools (Oakes, 2005).

Additionally, industrialization of the economy in the early 20th century, and the development of labor principles such as the Taylor System of scientific management began to reshape the American understanding of production. The Taylor System relied on "time-and-motion" studies to find the most efficient means of production, as well as the separation of planning from performance: management created the plans for work, while the workers carried them out. This system was expected to benefit both workers and owners of the shops. In the 1920s, new psychological testing developed by Hugo Münsterburg purported to also find the best workers for each job (Gillespie, 1991). Businessmen exposed to this system in factory production often sat on school boards and tried to apply similar principles to education. Early testing in math and spelling done by Joseph Mayer Rice found that many students were held back repeatedly until they eventually dropped out, and concluded that the school system was designed only for the best and the brightest, and was failing the rest of the students. Emphasizing a desire to create products (students) at low cost, schools began to move toward "within-school differentiation in the *kinds* of educational programs" (Oakes, 2005: p. 30) as a way to make schooling more efficient and utilitarian (Kliebard, 1986).

At first, students were openly classified by racial, ethnic and economic background, but with the popularization of IQ tests in the 1920s, assignment to track began to depend on "impartial" test results. Hugo Münsterburg's student, Robert Yerkes, oversaw the development of intelligence testing applied to US soldiers to help sort them into the "appropriate" roles (Gillespie, 1991). Although these tests were in fact heavily

culturally biased (about 80% of immigrants tested as "feeble-minded"), they were perceived in education as a way to identify and assign students so that they could be efficiently educated in the manner most conducive to their inherent ability level (Gould, 1996). Throughout the 1920s, many remained convinced that education could teach students to think and help individuals rise out of their station in society, but the popularization of psychological and IQ testing convinced others that intelligence was innate, and the purpose of schooling was to efficiently train children for the jobs they were predestined to hold. This led to the development of separate curricula for students in different walks of life (Kliebard, 1986). At the same time, the emergence of the school counseling movement added the aspect of student "choice" to track placement. However, student choice was also heavily influenced by circumstances, test scores and counselor and teacher expectations (Oakes, 2005).

In the 1930s, a movement began to change the purpose of schooling from efficiently educating different classes of students to preparing Americans to address social injustice. Spurred on by the market crash in 1929 and the subsequent deepening depression, followers of this movement argued that the quality of schooling mattered more than quantity or efficiency. The Progressive Education Association conducted an eight year experiment on curriculum at the school level, determining that the traditional college preparatory curriculum was not the only road to success in college. Some reformers even argued for increased integration in schools and educational programs, both racially, economically and across gender lines (Kliebard, 1986).

By the time the United States entered World War II in 1941, there were several distinct camps of education and curriculum reform, but the onset of war helped to

galvanize a focus on civic education, the sciences and math. In 1959 the Conant Report, named after former Harvard president James B. Conant, was issued in response to an increased interest in mathematics and science education spurred on by the Cold War and the launching of the Sputnik satellite in 1957. This report inspired reforms led by National Academy of Sciences, College Board, and eventually Congress. The authors of the report noted a lack of well-trained teachers and poor curriculum in math and science, and the reforms began with federally-funded production of curricular materials, which were eventually taken over by private producers. Only a minority of students was actually exposed to this material, and by the late 1970s, only high-track students used these materials. By 1985 American education was characterized by neither common core standards nor pure differentiation, and districts varied in the focus they put on different elements of curriculum (Resnick & Resnick, 1985).

In 1985 Jeannie Oakes's landmark book, *Keeping Track: How schools structure inequality* was released. In her study she found evidence of rigid differentiation by ability into separate courses of study (between which students found it difficult, if not impossible, to transfer) and a strong association between these track levels and race and socio-economic status. In addition, Oakes found qualitative evidence of lower instructional quality in the "low" tracks and worse achievement outcomes for these students. Several other studies released around this time found similar results (Esposito, 1973; Gamoran, 1987; Kulik & Kulik, 1982; Slavin, 1990), and these results, combined with growing public concern over racial and economic achievement gaps, put pressure on schools and districts to detrack in the name of equity. In the second edition of her book, Oakes found that rigid tracking systems were more or less abolished by the mid-1990s.

Since then most students have enrolled in high school courses separately, not as a part of a program of study.

Defining Tracking in a New Era. While rigid tracking across subjects has decreased, ability grouping at the classroom level remains pervasive. According to a study by Loveless (2009), the percent of schools with only *one* track has not increased in mathematics: about 85% of middle school students are still separated into at least two levels in their math classes. In the modern context of more flexible sorting systems, the language of tracking has become more complex, as researchers and policymakers continue to debate the significance and importance of sorting, with some arguing that less rigid systems are no longer detrimental to students because they do not have the impact of inflexible tracking (Lucas, 1999; Oakes, 2005). When applied to contemporary American public schools, therefore, "tracking" needs to be defined. For the remainder of this paper I will distinguish between "tracked" and "untracked" settings. "Tracked" settings have more than one level of mathematics at a given grade, so that students are either in an "advanced" or a "regular" course. "Untracked" settings do not group students in this way. Prior literature has used "detracked," "untracked," and "heterogeneously grouped" to refer to students who are not ability grouped at the classroom level, but I use "untracked" to emphasize two points. First, some untracked schools have never been tracked: historically they were too small or too homogeneous to sort students by achievement. Therefore, they have not been "detracked" in response to policy changes, but rather *remain* "untracked." Second, schools vary in the spread of achievement among their students, and so a tracking policy may not be the only predictor of the achievement heterogeneity within classes. For example, a student in a "high-track"

classroom in a school with a wide range in achievement may have more classmates of average achievement than a student in an "untracked" classroom in a uniformly high-achieving school. Thus, in this paper, the differences in achievement between "untracked" and "tracked" students are hypothesized to come largely from what takes place in the classroom (namely, the quality of instruction), rather than from a change in policy or the achievement levels of students' classmates.

Additionally, I will distinguish between "high-track" and "regular- or low-track" students. I divide the tracked students into two groups in part because of a finding by Loveless (2009) that even "untracked" schools often have two ability groups in middle school mathematics, but also because much of the debate over tracking remains a concern about the impact on high-achieving students. Analytically, it is difficult to find "low" track students labeled as such in American schools today; politically, there is a greater concern about the impact of detracking on the highest-achieving than on the lowest-achieving students.

Arguments for and against tracking. Many of the arguments for and against tracking can be categorized into debates of efficiency versus equity or equality (Hanushek & Woßmann, 2006). Proponents of tracking typically espouse an efficiency view: tracking can more efficiently target instruction to students, creating a greater output of exceptional students for fewer inputs of time, money and instruction. Opponents, meanwhile, argue that tracking is inequitable and unequal: it segregates students by race and socio-economic status and provides unequal education across the track levels, exacerbating existing inequalities in the United States.

The "efficiency" outcome, though not always stated in those terms, has generally been measured as the mean difference in achievement between tracked and untracked students. Proponents of tracking tend to focus on the achievement of the highest-ability students, arguing that detracking will harm them (Brookings Institution, 2009). Conversely, they also typically argue that tracking will increase the achievement of the high-ability without harming the achievement of low-ability students. Therefore, a key aspect of their argument is that average achievement under tracking is higher than average achievement in untracked settings.

The "equity" and "equality" outcomes have been measured in several ways. Many researchers focus only on the first concern: students are segregated by race and socio-economic status. They examine whether assignment to track levels is disproportionate by these variables (e.g., Braddock & Dawkins, 1993; Oakes, 2005; Rosenbaum, 1976), arguing that this indicates that student assignment to track levels is not "fair" or equitable, because it is based on factors other than ability. Other researchers have examined the gaps in achievement or achievement gains between high- and regular- or low-track students, arguing that growing gaps indicate growing inequality in our schools (Brewer, Rees & Argys, 1995; Gamoran, 1987).

Finally, some researchers have looked at how instruction differs between high- and regular- or low-track classes, arguing that tracking does not *target* instruction to the needs of students, which could be equitable, but rather perpetuates existing inequities by providing high-status knowledge only to students of high socio-economic status (Oakes, 2005). For example, Oakes (2005) found that students in high-track classes studied

literature and mathematics to prepare them for college, while students in low-track classes studied life skills, such as balancing a checkbook.

Research on Tracking Outcomes. A large body of research has been done to test both the efficiency and the equity/equality hypotheses. In 1982, Kulik and Kulik conducted a meta-analysis of existing research on the effects of ability grouping in secondary schools. They reviewed experimental studies of ability grouping in which students were grouped into classes by ability (as measured by IQ, reading tests or prior achievement). They found 38 studies in which they could calculate an effect size, but only ten of these studies showed statistically significant differences in achievement. Although eight of those ten favored tracking (higher achievement among tracked students than among untracked students), the average effect size was only 0.1 and the range was wide. Kulik and Kulik argued that this large range indicated that factors other than ability grouping were likely at play in these studies. Looking more closely, they found that studies examining gifted programs showed larger effects, while studies looking at programs for "academically deficient students" showed a near zero effect, as did studies looking at "unrestricted populations." These studies suggest that tracking policies that provide gifted programs may raise mean achievement (likely by raising the achievement only of those students), but tracking policies that isolate the low-achieving or that broadly apply to all students may have no overall effect.

Slavin's (1990) meta-analysis of tracking studies, which focused on comprehensive ability-grouping programs that "incorporated most or all students in the school" (p. 275), supports this conclusion. Slavin found a median effect size of -0.02 across the twenty studies that had computable effect sizes. Nine more studies found no

statistically significant effects, and when these were included, the median effect size of ability grouping was found to be zero, indicating no difference between tracked and untracked schools. Mosteller, Light and Sachs (1996) performed a similar review, adding a few studies from later in the 1970s and found the same result: an average effect of tracking that approaches zero. The real finding, according to these authors, was that there were not enough studies, the studies that existed did not use big enough samples or long enough time periods, and so the findings were not well established.

Overall, these meta-analyses find only small and often conflicting outcomes of studies comparing the average achievement of students in tracked to students in untracked settings. On average, it seems that tracking does not raise the overall achievement of students, indicating that it is not more "efficient," but it is still possible that tracking increases the achievement of "high ability" students while harming the achievement of "low ability" students. Thus, equal effects in opposite directions average out to no overall difference.

Research on the "equity" outcomes of tracking often focuses on the criteria for selecting students and whether the selection process results in segregation by race or socio-economic status. While the measures of ability used in tracking have varied, the justification behind them is commonly meritocratic selection: students with higher innate abilities receive more advanced instruction (Oakes, 2005). However, this argument rests on the assumption that the criteria are truly meritocratic: Rosenbaum (1976) argued that "permanent selection can be efficient only if they are based on a *valid* and *stable* criterion … and *completely based on the criterion*" (italics in the original, p. 52). In examining the measures used in one town in the Boston area, he found that measures of the three types

of criteria used (ability, effort, and achievement) varied in validity, but none were stable. Additionally, although there was a strong correlation between all the indicators and track placement, none of them was a perfect predictor, even at the extremes. Rosenbaum argued that the use of invalid and unstable indictors to place students in track levels undermines the efficiency argument in favor of tracking: to target instruction efficiently, students must be accurately sorted by ability. However, the use of invalid and unstable indicators for placing students can also affect the equity of tracking by making the process less meritocratic.

Finally, Rosenbaum argued that, although students had some choice in their track placement, in practice this was not an informed or a free choice. Although most students made the college track their first choice, they were often influenced by guidance counselors to choose lower tracks, to switch to lower tracks or to put off switching to higher tracks throughout their time in school. These findings indicate that track placement may not be as "meritocratic" as its supporters claim it to be.

Later studies have also found that African American and Latino students are significantly over-represented in low-track mathematics and English classes (Braddock & Dawkins, 1993; Spielhagen, 2006). On the other hand, some studies find that race and socio-economic status are no longer significant predictors of track placement and the middle and high school levels when controlling for prior achievement, (Archbald, Glutting & Qian, 2009; Schmidt, 2011). Therefore, this over-representation of minorities in lower track classes may be caused in part by their prior achievement, particularly by the time students reach middle school. In other words, although students may be placed

primarily based on prior achievement, this results in disproportionate representation of minorities in low-track classes because of existing achievement gaps.

Nonetheless, this disproportionate placement may exacerbate those existing achievement gaps if track level is associated with differential achievement gains. Using student-reported track level in the High School and Beyond dataset, Gamoran (1987) found that high-track students significantly out-performed other students, even when controlling for background (gender, socio-economic status, race) and prior achievement, and this effect was largest in math. Additionally the gap between academic and vocational/general track was three times the size of the gap between vocational track and dropouts. Brewer, Rees and Argys (1995) also found that students in low-track classes had significantly *lower* achievement, and those in high-track classes had significantly *higher* achievement than heterogeneously-grouped classes, and the impact was about the same size in each direction.

Finally, Hanushek and Woßmann (2006) examined differences in tracking and achievement across twenty-six countries, including the U.S., Canada, Germany and France. This study used cross-country comparison, comparing tracked to untracked settings within each country to account for existing differences between countries. The authors find that "Relative inequality increases in every country with tracking except the Slovak Republic, while relative inequality decreases in every country without tracking except for Sweden and Latvia" (p. C69). Controlling for existing differences in the level of inequality, countries that track students before the age of 15 had larger inequality in secondary school on all three measures used.

Overall then, studies suggest that tracking has little effect on the mean achievement of students, but it may be inequitable, as it is not always based on valid and reliable measures of ability, and it increases inequality by raising the achievement of the already high-achieving and depressing the achievement of low-achieving students. In the United States this may be especially problematic because of the association among race, socio-economic status, and track level.

Instruction as the mechanism. Where the efficiency and equity/equality schools of thought on tracking share common ground is the belief that instruction is a key to success and failure. These arguments rest on the assumption that teachers and instruction can have an important impact on student achievement (Darling-Hammond, 1998; Lucas, 1999). The "efficiency" perspective argues that tracking will improve achievement overall because teachers will be able to more efficiently target their instruction if they have classes that are homogeneous by ability (Brookings Institution, 2009). The "equity" camp argues that this "targeting" of instruction is, in fact, rationing of high-status knowledge (types of knowledge and ways of thinking that are valued by the upper classes) so that only those who are already of high status may have access to it, and this rationing will perpetuate existing inequalities, particularly racial and socio-economic inequality (Darling-Hammond, 1998; Oakes, 2005). In both cases, instructional quality can be characterized as a "mediator," or the mechanism through which tracking policy affects achievement (Baron & Kenny, 1986).

Although instruction forms the cornerstone of both these arguments, quantitative findings to support these points, particularly in mathematics, are slim. The most important study in this area is Jeannie Oakes's 1985 book, *Keeping Track,* mentioned

above, which found that low-track classes focused more on life skills and reading, writing and math for employment and everyday life, and the amount of time spent on academic tasks was less in low-track classes. High-track students were exposed to more "high status knowledge" of the type required for college. Oakes did not address the relationship between the *quality* of instruction and tracking, however, as at the time she found no strong links between practices and outcomes in prior research. While to some extent this remains true today, there have been more studies providing suggestive evidence of the link between particular instructional practices and outcomes in mathematics since 1985, as I describe below.

Throughout the 1980s, research on the role of the teacher and instruction in the relationship between tracking and student achievement often focused on the elementary grades, and particularly on ability grouping in reading. These studies found that ability group level affected teacher behaviors such as behavior management (Eder, 1981), the number of words taught (Dreeben & Barr, 1988; Gamoran, 1986), and the amount of time spent on instruction (Dreeben & Barr, 1988). These studies also connected ability grouping to outcomes such as the number of words students learned (Dreeben & Barr, 1988; Gamoran, 1986) and student attentiveness (Eder, 1981). What these studies did not explicitly measure was how instruction may intervene in the relationship between tracking and achievement outcomes (the mediational impact of instruction). Instead, they modeled the impact of tracking on instruction and the impact of instruction on outcomes separately.

When extended to middle and high-school grades, most quantitative research has focused on reading and English Language Arts (ELA) rather than mathematics.

Although covering a different subject area, some of these studies provide a hint at how to measure the relationship between tracking, instruction and achievement. Beginning with practices shown by prior studies to lead to higher achievement, such as coherent discussion, revising activities and homework, these studies found that such practices were less frequent in low-track classrooms than in high-track classrooms (Applebee, Langer, Nystrand & Gamoran, 2003; Gamoran, Nystrand, Berends & LePore, 1995). Several of the studies also used methods such as multi-level modeling and mediational models to show that 1) these practices were linked to achievement in their data, and 2) these practices mediated the relationship between tracking and achievement (Applebee et al., 2003; Carbonaro & Gamoran, 2002). Using data from the National Education Longitudinal Study of 1988 (NELS:88), Carbonaro and Gamoran (2002) found that differences in teacher-reported instructional practices and content accounted for 10 to 20% of the tracking effect and about 30% the socio-economic status gap between students.

In mathematics, very few studies have used the approach deployed in English Language Arts (starting with proven instructional practices and linking them quantitatively to the impact of tracking on student achievement). Instead, the vast majority of mathematics instructional studies have been small, qualitative examinations of teacher practice in a few classrooms or schools. For example, Boaler and Staples (2008) compared one or two teachers in each of three schools (dubbed "Railside," "Hilldale" and "Greentop") using classroom observation, student interviews, questionnaires and student achievement. The authors found that students in the two tracked schools, Greentop and Hilldale, spent more time in lecture, and the teachers

posed shorter and less conceptual problems to students than in Railside, the untracked school. At Railside, teachers asked more varied questions, students reported enjoying math more, and by the end of two years, they significantly outperformed the other schools' students on a test in algebra and geometry, even though their achievement began lower than students in the other schools. However, this analysis did not account for other pre-existing differences between schools or students, which may lead to omitted variable bias.

Reed (2008) conducted interviews and classroom observations with two National Board Certified mathematics teachers, both teaching two different track levels of mathematics. Reed's teacher-subjects felt they had to make the material in regular classes less rigorous than that in honors classes to deal with the wider range of skill levels. Additionally, teachers used more direct instruction and less group work, shortened tasks and used more scaffolding and reviewing of old concepts in the "regular" classes. Additionally, teachers more often did the math themselves in front of the regular classes as compared to the honors classes. In honors classes, teachers usually introduced the task quickly and then had students work alone or in groups; while in regular classes they walked through problems step by step with the class.

In both of these studies, the authors found qualitatively important differences in instruction that seemed to be linked to tracking, such as the types of tasks used and how those tasks are introduced to students. However, they did not quantitatively examine the size of those differences, nor did they statistically control for pre-existing differences between the schools and classes. Although Boaler and Staples (2008) also examined

students' achievement outcomes, they did not link those outcomes to either the instructional differences or to the tracking in the schools.

In a similar study, Horn (2006) examined successful detracking initiatives in two mathematics departments, one in England and one in the United States. Both schools detracked by placing all students in college preparatory mathematics in ninth grade. From the case studies on these two schools, the author concluded that schools that successfully detrack math share four characteristics: 1) a "connected and meaningful view" (p. 73) of math, including connections to the real world; 2) a focus on important mathematical ideas, instead of just practicing procedures), 3) a balance of coordination and professional discretion in teaching decisions (e.g., coordination across grade levels); and 4) "clear distinctions between *doing math* and *doing school*" (p. 78, e.g., challenging assumptions about who is "smart" and good at math using group work). Although these conclusions align with the vision of mathematics instruction supported by the NCTM, as will be shown below, quantitative differences between tracked and untracked schools on these characteristics were not tested by Horn.

Gamoran and Weinstein (1998) conducted one of the largest case studies on the relationship between instruction and tracking in mathematics, interviewing teachers, principals, district personnel and parents and observing mathematics and social studies classrooms in 24 restructured schools across the country. Classroom instruction was rated for "extent of higher-order thinking, depth of knowledge, and substantive conversation" (p. 389). Although all high schools were concerned with providing high quality instruction to diverse learners, most schools were not successful. The study found evidence that in tracked schools, higher quality instruction was found in higher track

classes, while in untracked schools, the teachers struggled with teaching mixed ability classrooms and often lowered the overall rigor of instruction. Only one high school was successfully providing a rigorous curriculum to classrooms of mixed ability students. Still, these findings were all qualitative: the authors found no statistically significant differences in the ratings of instructional quality by tracking or track level in high school. Additionally, as in the studies above, Gamoran and Weinstein did not link differences in instructional quality to student achievement. On the other hand, unlike many other qualitative studies, it did begin with a quantitative measure of high quality instruction, rather than using an inductive method and allowing categorical differences to arise from the data.

What makes "good" mathematics instruction? The qualitative studies discussed above often varied in their definition of "good instruction." One reason may be that they varied in the contexts they examined, their methods, and the subjects of interest. To answer the question of whether instructional quality has an impact on the relationship between tracking and achievement, we must start with a mathematics-specific definition of high quality instruction. The definition I will use begins with the National Council of Teachers of Mathematics (NCTM) standards, which are cited by many as what counts as "good" mathematics teaching and learning (e.g., Freeman and Crawford, 2008). These standards focus on the teacher knowing "what students know and need to learn," challenging all students and emphasizing conceptual understanding over procedural fluency alone (NCTM, 2000). Conceptual understanding here refers to a student's ability to grasp the mathematics on a deeper level and flexibly and independently apply it in new situations, rather than carry out a set of memorized steps to solve a problem. The focus

on conceptual understanding connects to prior research by Oakes (2005) and others discussed above, who argued that a focus on explanation, justification and deeper understanding, in contrast to simply carrying out known procedures, prepares students for higher status jobs where more independent thinking is required. Opponents of tracking who make the equity argument contend that the tracking system may perpetuate inequalities between students by reserving this type of high quality instruction for the higher track students. Hence, differences in the quality of instruction between tracks may be the mechanism by which the policy affects student achievement. I will apply this dichotomy between 1) instruction aimed at supporting students to develop conceptual understanding of key mathematical ideas *and* procedural fluency and 2) instruction aimed *only* at supporting students to develop procedural fluency to three components that have been the focus of prior research on the quality of instruction in mathematics in particular: the cognitive demand of the task posed to students (task potential), whether the cognitive demand is maintained over the course of the lesson (task implementation), and the quality of a concluding whole-class discussion focused on students' solution to a task.

Several researchers have found a link between the cognitive demand of the task (mathematics problem or assignment) that the teacher provides to the student and their subsequent learning **(**e.g., Boston and Wolf, 2006**)**. These researchers have argued that high-quality tasks emphasize conceptual understanding through the use of multiple representations, multiple solution paths and multiple entry points; they are relevant to students' lives, require problem-solving and emphasize meaning and reasoning rather than basic skills (Horn, 2006; Rubin, 2006; Wheelock, 1992).

Stein, Grover, and Henningsen (1996) outlined a hierarchy of tasks that is aligned with the above components. Low-cognitive demand tasks are those that require students to carry out routines without any connection to the underlying mathematics, called "procedural tasks". One example of a procedural task may be a worksheet of problems using the Pythagorean Theorem where students are asked to apply the formula to a series of triangles. High-cognitive demand tasks are those that stress justification, reasoning and making connections, in addition to learning the required procedures. They call this "doing mathematics." An example of a high cognitive demand task is one that asks students to come up with an equation representing a complex pattern represented in the problem and then explain why their equation works. Stein, Grover, and Henningsen (1996) observed eight teachers in four schools and found that the majority of teachers introduced tasks that included multiple solution strategies and/or multiple representations of the problem, often falling into the "doing mathematics" category. Despite choosing high-level tasks, teachers were unlikely to maintain the cognitive demand throughout the remainder of the lesson. In fact, the higher the cognitive demand of the task, the less likely that the difficulty would be maintained.

In addition to selecting a high-quality, demanding task for students, teachers must ensure that the task is actually implemented at this same high level in their classroom. While cognitively demanding tasks are important to building students' conceptual understanding, to reach this goal students must actually engage in the difficult aspects of the problem. Stein, Grover and Henningsen (1996) found that when teachers introduced tasks emphasizing procedures *without* connections to the underlying mathematics, 96% of these were maintained at that level. Conversely, when tasks emphasized procedures

with connections, 53% of them declined in cognitive demand. Finally, tasks emphasizing "doing mathematics" declined 62% of the time (p. 478). For example, teachers often selected tasks that called for explanation or justification of students' solutions, but did not follow through on requiring these explanations.

Henningsen and Stein (1997) also examined how mathematical tasks with high level cognitive demand may end up implemented as more low-level tasks once students begin working on them on their own or in groups. Using data from classroom observations on twelve teachers in four schools, the authors isolated 58 of 144 tasks that had a high level of potential as written. Teachers maintained the cognitive demand of those high-level tasks in only 22 cases, while in 36 cases they decreased the rigor of the task. In classrooms where the cognitive demand was maintained at a high level, common teacher behaviors included building on prior knowledge, using scaffolding, providing sufficient time for students to work together, modeling high-level performance for students and pressing for explanation and justification. The teachers who decreased the demand of tasks commonly shifted the emphasis of the task away from the meaning behind it and toward the correctness of students' answers. Teachers may also have taken away the parts of the lesson that were challenging, either by providing students with the "correct" procedure or doing the procedure for them.

While Henningsen and Stein (1997) focused on the ways the teacher may reduce the cognitive demand of tasks when s/he introduces the lesson to the students, Cohen's (1972) book *Designing Groupwork* examines how tasks play out once students are allowed to work on them in groups. In this book Cohen defines groupwork as "students working together in a group small enough so that everyone can participate on a task that

has been clearly assigned… without direct and immediate supervision of the teacher" (p. 1-2). Although the teacher is not immediately supervising, successful groupwork depends in several key ways on teacher behaviors, including delegating authority to the students and providing tasks that cannot be completed by an individual student. Research has shown that groupwork is effective for conceptual learning, helping students understand concepts by explaining them to others or having them explained by a peer. To be effective the tasks must be conceptual, the students must actually talk to one another, the group must have the resources required to be successful, and the teacher must attend to status issues, cooperation and other norms of participation. Cohen argues that groupwork can also be useful for building academic social skills, increasing time on task and managing a variety of incoming achievement levels. Groupwork addresses incoming achievement as long as the groups are heterogeneous and students are trained to cooperate, to recognize when others are struggling, and to use one another as a resource.

The ability and willingness of teachers to implement rigorous tasks at their intended level, both during the set up and during groupwork may be particularly important in untracked settings. Boaler (2006) found that that detracking leads to more equitable outcomes, but only when paired with certain practices to make grouping more effective, including using group-worthy problems (open-ended, with multiple solution paths). Likewise, Cohen and Lotan (1997) found that cooperative learning had a strong and significant relationship with achievement in heterogeneous classrooms. They argued that other studies did not find this result because the teachers did not use group-worthy tasks and other complex instruction techniques. Therefore, task implementation as discussed above could be a mediator between tracking policies and student outcomes.

A third component of high-quality mathematics instruction that has been a focus of prior research is classroom discussion. During classroom discussion, students have the opportunity to engage in what Cobb, Boufi, McClain, and Whitenack (1997) call "reflective discourse." Reflective discourse refers to a process in which the teacher and students work on a problem and then reflect on the work they have done as a focus of discussion, coming to a deeper or new understanding. In mathematics classrooms, teachers introduce a task, provide students with time to work on the task alone or in groups, and then bring the class back together to discuss the task as a whole group. In Cobb et al.'s approach, the "collective reflection" in this final discussion creates the conditions necessary for students to come to a deeper understanding because the process of learning math is inherently social. In many cases, this final whole-class discussion does not take place, or if it does teachers may ask students to share their solutions but not press for justifications or explanations of why their solution works, and they may not make connections between different solutions. Classrooms that use less discussion or have discussions that do not push for abstraction and reflection prohibit optimal student learning. These practices are also important because they are used to deepen students' conceptual understanding and provide access to the high status knowledge that is hypothesized to be important to both learning and preparation for higher status jobs, which require the ability to reason and justify one's answers, as opposed to being able to carry out a known procedure. Therefore, if lower track classrooms have lower achievement gains, it may be due to differential use of mathematical discussion.

Research Questions

Based on the definition of high quality instruction outlined above, I address the following research questions: 1) Are there measurable differences in student achievement by track level? 2) Are there measurable differences in instructional quality between teachers in tracked and untracked settings? 3) Between high- and low-track classrooms? 4) Is rigorous mathematics instruction associated with higher student achievement? 5) Do differences in instructional quality mediate the relationship between track level (high versus regular track) and student achievement?

Data and Measures

To answer these questions I use four years of data from the Middle school mathematics in the Institutional Setting of Teaching (MIST) project at Vanderbilt University. MIST is a National Science Foundation-funded project that examined the relationship between institutional supports, instructional practices and student achievement in 30 middle schools in four large, urban districts between the 2007-08 and 2010-2011 school years. These districts were selected because they were undertaking instructional improvement initiatives in mathematics that were aligned with the NCTM standards and goals for student learning and had adopted inquiry-oriented curricula, such as the Connected Math Project (CMP). In each of the four districts, six to ten middle schools were selected in collaboration with central office staff to be representative of the district. Within these schools, teachers were recruited to participate in the study. Between 17 and 38 teachers participated in each district in each year. Because participating teachers left the study and new participants were added, this amounted to 223 unique teachers and 9,847 students in the observed classrooms. Over 100 of these

teachers were only observed in one year, while 39 were observed in two years, 44 were observed in three years, and 33 teachers were observed every year for four years.

These teachers were videotaped during instruction for two consecutive days between January and March. As a part of their participation in the study, they were asked to do a problem-solving lesson with related whole-class discussion on the days of filming. The students' achievement data were also collected, including the current year mathematics achievement test results as well as two prior years' scores, which were standardized to district averages and standard deviations by grade and year. Although nominally collected, the test score from two years prior was missing for the majority of students. Therefore, only one prior year of achievement is used in analysis. Student demographic data were also collected, including grade level, gender, race, free/reduced-price lunch, English Language Learner and Special Education status.

There are three main measures required to answer my research questions: tracking measures, measures of instructional quality, and measures of student achievement.

Tracking variables. Tracking variables were created by examining the class-level data shared by the districts (course name and track level if specified), and filling in with teacher and principal reports from one-on-one interviews where necessary. In these interviews, we asked whether classes were grouped by skill level (tracked) and what those levels were (track level). Principals provided an overall view of the courses offered at the school and how students were placed in them, while teachers provided the information on their particular classes. We then used this this information to verify

course files from the school and district. If a track level still could not be determined, we used the average prior achievement of the students in the course to assign a track level[1]. From this information we created two variables. First, "tracked" indicates whether students in that grade level in that school were grouped by ability. If one course in a grade was tracked (8th grade Algebra I, for example), we labeled the entire grade as tracked, because students were chosen for that higher-level course while other students were chosen for the "regular" level course. However, we did not label the entire *school* as tracked, because some schools did not group students into ability levels in mathematics until 8th grade.

Second, within tracked grades, the "high-track" variable indicates whether the individual student was in a "high-ability" course. High track classes were courses such as honors, advanced, Advanced Placement, and high school Algebra I courses offered in eighth grade. Although high track courses were usually easy to identify, it proved much more difficult to distinguish "low" track courses from "regular" track courses. It was very rare that the course data provided by the districts would identify a "remedial" or below grade level course, with the exception of "double dose" courses: a full second period of mathematics offered in addition to a regular math course. These types of courses were not videotaped and thus could not be used in this analysis. Due to this difficulty in separating low track courses from regular track courses, we created one group called regular/low track. This resulted in three categories for courses: untracked, regular/low track and high track. Throughout the rest of this analysis, I will refer to

[1] We primarily used this information to triangulate with the teacher and principal interviews, but it was also employed to fill in gaps if they did not respond to the question, or the response was not clear.

regular/low track as "regular track," as there were few identifiable "low track" classes in this group.

Instructional Quality. To measure instructional quality, MIST used videotapes of teachers' classroom instruction to score teachers on the Instructional Quality Assessment (IQA) each year. Boston (2012) outlined the development of this set of rubrics used to measure instructional quality. Prior attempts at measuring instruction included student and teacher self-report, which are prone to error and bias, as teachers may misremember or misrepresent their actual practices. Another common failing of previous rubrics was a lack of focus on "a specific theory of instructional practice" (Junker et al., 2006, p. 3), making the resulting ratings subject to rater unreliability and a lack of internal coherence. The IQA is used to rate live or videotaped classroom instruction, and it was developed to map onto specific Principles of Learning developed by the Institute for Learning at the University of Pittsburgh. The four principles are: Academic Rigor, Clear Expectations, Self-Management of Learning, and Accountable Talk. The authors developed the IQA rubrics to be specific, measurable and "low inference." This makes it possible to train raters who do not have extensive content knowledge without resulting in significant rater effects. In a pilot study in sixteen elementary school classrooms in two districts, the authors found that they had high inter-rater reliability and were able to distinguish significant differences between districts in instructional quality. A newer version of the IQA rubrics were used to rate the instructional quality of teachers' classrooms as a part of the Middle-school Mathematics and the Institutional Setting of Teaching (MIST) project. They were chosen because of

their alignment with the type of instruction supported by the National Council of Teachers of Mathematics as well as the high reliability and validity of the measures.

This study focuses on IQA's Academic rigor rubrics: Task Potential, Implementation, and Discussion.[2] The Task Potential rubric is based on ratings of the task *as written*. As discussed above, high cognitive demand tasks are conceptualized as those with multiple solution paths and those that allow students to make connections between ideas and communicate their thinking[3]. For example, a task may provide students with a list of supplies and their price and ask them to build a fence around an uneven farm property on a certain budget. To be considered high level, it must also require the students to discuss and justify their choices and explain why their solution meets the requirements of the problem. A low-level task, on the other hand, may ask only for the formula for the area of a trapezoid or ask students to find the area of a series of shapes.

The Implementation rubric rates the level of cognitive demand *actually* required during the class period. The authors of the rubric (Junker et al., 2006) argued that teachers can and often do change the level of cognitive demand of a task over the course of the period, and students learn best when the task is high quality *and* the rigor of the task is maintained throughout the course period. For example a teacher may provide students with the farm task above, but provide them with the steps to dividing the property and the amount of materials required for each shape, explaining to them how to

---

[2] The original IQA rubrics also included a "Teacher Expectations" rubric, but this rubric was not applied to the MIST data because of the increased demand on teachers and so is not included in this analysis.
[3] The levels of this rubric are derived from Stein, Grover and Henningsen (1996).

solve the problem. As a result, the task that students engage in would be closer to the low-level worksheet task than the high-potential task as it was written.

Finally, the Discussion rubric "provides an overall, holistic rating of the level of cognitive processes evident during the final discussion of the lesson" (Boston, 2012, p. 84), based on how students are encouraged to explain their thinking to the teacher and to one another. High quality discussions require students to justify and compare their solution strategies to come to a deeper understanding of the underlying mathematics in the problem. Again, in the problem discussed above, the teacher may ask students to come together at the end of class and share their solutions and their justifications, making connections between different strategies and helping build a deeper understanding. Conversely, the teacher may have each group share their solutions without discussing how they made decisions or found their answers.

As designed, each of these rubrics is on a four-point scale, and the ratings are comparable across dimensions. Levels three and four correspond to high-level cognitive demand occurring in the classroom, while levels one and two show low-level cognitive demand. The Discussion rubric also includes a zero score, signifying that no concluding whole-class discussion was held. In addition to these three rubrics, a combined IQA score was created. This combined IQA score averages the discussion sub-scores with the academic rigor sub-scores, giving more weight to specific "accountable talk" moves, for a combined IQA score. Although each teacher was observed twice per year, their best score of the two days was taken as their IQA score for this analysis. Therefore, there is only one IQA score on each rubric per teacher per year.

In 2006, Resnick, Matsumura, and Junker assessed the reliability and validity of the IQA. The authors applied the IQA to observations of thirteen 6th and 7th grade math teachers in five urban middle schools in one district. Teachers were rated on the IQA rubrics in two consecutive lessons. In this pilot, inter-rater reliability was very high: 81.8% overall and ranging from 70% to 100% on each rubric. The authors also found "as few as two observations yielded a stable estimate of quality, when teachers complied with the requirements of the data collection" (p. 17). Teachers were asked to have a lesson with a discussion component on the days of observation, though several teachers did not do so, resulting in zero scores on all the Accountable Talk rubrics for those days. This reduced the stability of measure when those lessons were included. Excluding those teachers, the authors found a significant relationship between IQA scores and student achievement, controlling for demographics and prior achievement. An increase of one IQA point was associated with a predicted increase of 0.16 standard deviations in achievement on the total math score and 0.32 standard deviations on the "procedures" subscale of the math test. I will examine whether this relationship is supported in the MIST data in answering the third research question. Overall, the study found that instructional quality varied a great deal, but that the average quality was "basic" (scores of 1 or 2 on a four-point scale).

Student Achievement. The final measure used in this analysis is student achievement, reflected in the state-mandated end-of-year assessment for each district. The districts provided scale scores for students as well as the distribution of scores at the state level. Because the four districts are in three different states, the same test was not administered to all students. For this reason, students' scores were z-scored to the state

distribution, so that each student's score is given in terms of its distance in standard deviations from the state mean. This allows for a comparison between students in different districts, despite a difference in the scale of the tests. As will be shown below, the average achievement in these schools and districts was below the state average achievement, so the majority of students have negative z-scores. In addition to this, I will use district fixed effects in all my models, to account for additional differences in the content tested on each assessment.

Methods

The data from the MIST project allows me to address my research questions using a large dataset and quantitative rather than qualitative methods. The first research question uses the MIST dataset to establish the relationship between track level and student achievement that has been found in prior research. For this model, student achievement is the outcome, and this analysis largely serves to establish the relationship I am attempting to explain using instructional quality as a mediator. As mentioned above, the dataset used for this paper includes students of selected teachers in selected classrooms and schools in four large, urban districts. This clustering of the data suggests a multi-level model approach to avoid the problem of correlated error terms (Raudenbush and Bryk, 2002)[4]. As each teacher is observed in only one classroom per year, but can

[4] Each teacher was observed twice in the same classroom, but never across classrooms in the same year. Therefore, I did not use teacher fixed effects to compare instruction across track levels within teacher in the main body of the analysis because it requires comparing the same teacher across years. However, the models controlled for classroom characteristics (racial concentration, percent free/reduced-price lunch, percent limited English proficient, class size and grade level) and student characteristics (race, prior achievement, free/reduced-price lunch status, limited English proficient and special education status). Additionally, teacher and school fixed effects were explored as a sensitivity analysis.

be observed more than once over the course of the study, these models have five potential levels: students nested within observations (a classroom in a year) nested within teachers nested within schools nested within districts. However, I use district fixed effects instead of introducing a district level to account for differences in the tests and contexts. Therefore, I tested unconditional four-level models. I found that, controlling for district, 8.8% of the variation in math scores was at the school level, 5.2% was at the participant/teacher level, and 18.3% was at the observation/classroom level. This is significant variation at all four levels, indicating significant dependency in the data.

Therefore, the model for this research question has four levels: the students at level 1, the year of the observation at level 2, the teacher at level 3 and the school at level 4. This is shown in Equation (1):

(1)

$$\textbf{Level 1: } Achievement_{ijkl} = \pi_{0jkl} + \pi_{1jkl}S_{ijkl} + e_{ijkl}$$
$$\textbf{Level 2: } \pi_{0jkl} = \beta_{00kl} + \beta_{01kl}High\ Track_{jkl} + \beta_{02kl}Z_{jkl} + \beta_{03kl}Yr_{jkl} + e_{0jkl}$$
$$\textbf{Level 3: } \beta_{00kl} = \gamma_{000l} + r_{00kl}$$
$$\textbf{Level 4: } \gamma_{000l} = \delta_{0000} + \delta_{001l}D_l + u_{000l}$$

Achievement$_{ijkl}$ is the achievement of student *i* in year *j* with teacher *k* in school *l,* S_{ijkl} is a vector of student controls (race, free/reduced-price lunch, special education status, limited English proficient status, prior achievement), Z_{jkl} is a vector of classroom controls specific to that year's observation of that teacher (grade level, percent free/reduced-price lunch, percent LEP, percent special education, percent minority and number of students), Yr_{jkl} is the study year in which the observation took place, and D_l is

the district fixed effect[5]. "High Track" is the track level variable, which appears at the observation/year level, because a teacher may have a high track class observed in one year and a regular or untracked class observed in the next year.

Research Questions 2 and 3 seek to establish the quantitative relationship between tracking and instructional quality. To address these questions, the dependent variables are IQA ratings on the "Task Potential," "Implementation," and "Discussion" rubrics, as well as a combined IQA measure. Although the IQA scores are on an ordinal scale, I dichotomize them to compare high quality (levels 3 and 4) to low quality (levels 1 and 2) instruction. The split between IQA level two and three is supported by the literature in the focus on the difference between instruction that emphasizes procedural learning without connections to the underlying mathematics and instruction that emphasizes conceptual learning (e.g., Horn, 2006; NCTM, 2000; Oakes, 2005; Stein et. al, 1996).

There are potentially four levels to the data used to answer the second two research questions: observations (classrooms in a particular year) nested within teachers nested within schools nested within districts. There is significant variation in IQA scores by district, but as in Equation (1), I use district fixed effects instead of a district level. To establish whether the other three levels are necessary, I examined the variation in IQA scores by school, participant and observation. Within district, there is little variation *between* school means in IQA scores, except in District A combined IQA scores. This is shown in *Figure 1*, where each bar represents the distribution of IQA scores in a school.

[5] I did not include school control variables in any of these models because they prevented the models from converging, indicating that the model was not identified. Instead, I examined school fixed effects to account for all time-invariant differences between schools.

Most of the variation appears to be *within* school, either between or within participants. If we examine variation by participant (see *Figure 2*), most of the variation seems to be within participant, but there is variation between participants as well.

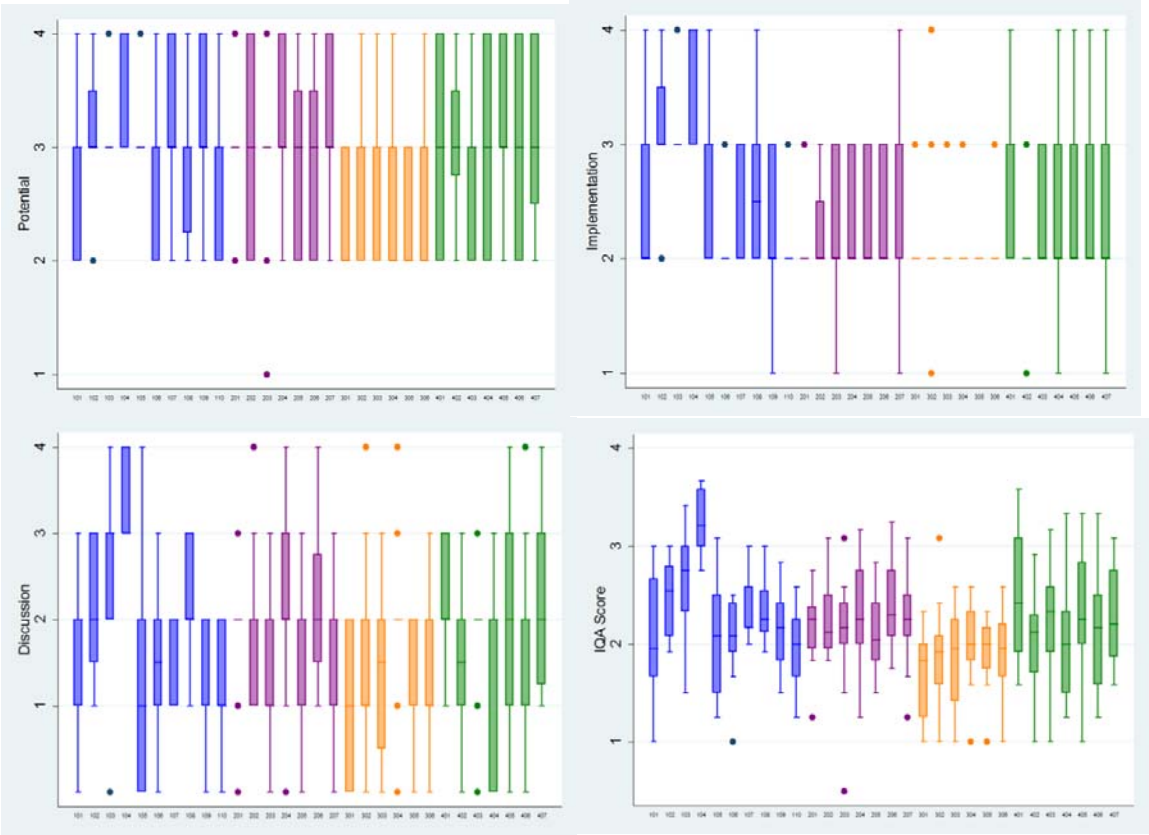

Figure 1:
Box Plots Representing the Distribution of IQA Sub-scores by School and District

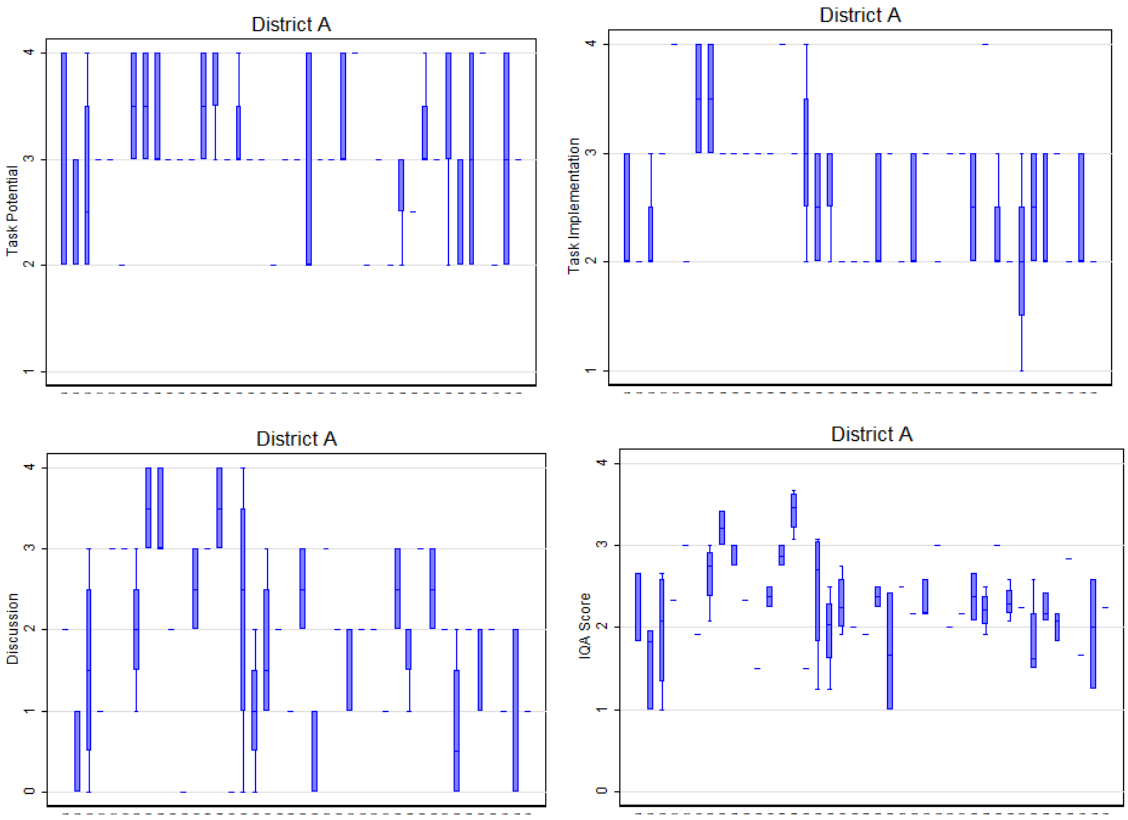

Figure 2:
Box Plots Representing the Distribution of IQA Sub-Scores by Participant Teacher and District

I tested unconditional models with three levels (observation, participant and school) for each of the IQA variables, and found that, controlling for district, less than 1% of the total variation in Potential, Implementation and Discussion scores was *between* schools. Nearly 20% of the variation in Implementation, 32.5% of the variation in Discussion and 17.9% of the variation in IQA combined scores was between participants. These findings suggest two level models for Implementation, Discussion and IQA overall: observation/year nested within teachers. The dependency at the participant level is less in Task Potential scores, but for consistency I use two levels for all four models. I also allowed the slope between IQA scores and study year to vary randomly between

participants, but this was not statistically significant for any of the IQA variables. Although there was significant variation in IQA scores by year, this variation seems to be similar across participants. Therefore, I fixed the slope between time and IQA scores at the participant level.

Therefore, the models for the second two research questions have two levels: observations nested within teachers. The tracking variables are observation-level variables, as the teacher may teach a high track classroom in one year (observation) and a regular track class in another year. Likewise, the teacher's instructional quality or IQA scores are at the observation level. I also control for other classroom characteristics, particularly the percent free/reduced-price lunch, percent minority, class size, percent Limited English Proficient and percent Special Education. Finally, I include district and study year fixed effects. The study year fixed effect is not to be confused with the observation/year level. While each teacher can have up to four observations, the study year in which these took place may vary. In other words, one teacher may have been observed in years 2 and 3, while another was observed in years 3 and 4. Both teachers would have two observations at the observation/year level, but the value of study year would be different. If there is a systematic relationship between IQA scores, study year and track, this would bias my results unless I controlled for year. Therefore, the models for the second analysis are of the form:

(2)

$$\textbf{Level 1}: IQA^{*}_{jk} = \beta_{0k} + \beta_{1k}Tracked_{jk} + \beta_{2k}Z_{jk} + \beta_{3k}Yr_{jk} + r_{jk}$$
$$\textbf{Level 2}: \beta_{0k} = \gamma_{00} + \gamma_{10}D_{k} + u_{0k}$$

There are four iterations of Equation (2), one for each of the IQA rubrics (Task Potential, Implementation and Discussion) and one for the IQA combined score. In this

equation IQA_{jk}* represents the latent, or underlying, construct measured by each IQA rubric as measured in classroom *j* in school *k*, Z_{jk} is a vector of classroom controls, Yr_{jk} is the year of observation *j* with teacher *k,* and D_k is the district fixed effect [6]. Using multi-level logistic regression models, I estimate the marginal probabilities of having a high IQA (a score of three or higher) for teachers in tracked and untracked settings, holding other factors constant. Similar models are employed to answer the third research question, comparing teachers in high- and regular-track classrooms, as shown in Equation (3).

(3)

$$\textbf{Level 1}: IQA^{*}_{jk} = \beta_{0k} + \beta_{1k} HighTrack_{jk} + \beta_{2k} Z_{jk} + \beta_{3k} Yr_{jk} + r_{jk}$$
$$\textbf{Level 2}: \beta_{0k} = \gamma_{00} + \gamma_{10} D_k + u_{0k}$$

My fourth research question examines whether the measure of instructional quality being used (the IQA) is associated with increased mathematical achievement on the high-stakes state tests. While there is a great deal of theory associating rigorous mathematics instruction of the type measured by the IQA with increased math learning, and prior research by the developers of the IQA found and association with student achievement, the relationship to these particular tests in these districts must be established. If IQA-type instructional quality is not associated with math learning, or if it is associated with math learning of a type that is not reflected on these tests, then it is

[6] I will not include teacher controls because I am not interested in explaining why a teacher's instructional quality might be higher or lower, but only if teachers in tracked classrooms have different instructional quality than those in untracked classrooms.

unlikely to mediate the relationship between tracking and student achievement on the tests.

For these models, student achievement is the outcome, therefore I use a four level structure, as in research question 1: students nested within observations nested within participants nested within schools. The models to answer this question are of the form: (4)

$$\textbf{Level 1}: Achievement_{ijkl} = \pi_{0jkl} + \pi_{1jkl}S_{ijkl} + e_{ijkl}$$
$$\textbf{Level 2}: \pi_{0jkl} = \beta_{00kl} + \beta_{01kl}IQA_{jkl} + \beta_{02kl}Z_{jkl} + \beta_{03kl}Yr_{jkl} + e_{0jkl}$$
$$\textbf{Level 3}: \beta_{00kl} = \gamma_{000l} + r_{00kl}$$
$$\textbf{Level 4}: \gamma_{000l} = \delta_{0000} + \beta_{0001}D + u_{000l}$$

S_{ijkl} represents a vector of student-level controls (prior achievement, race, free/reduced-price lunch status, special education, and limited English proficiency), and Z_{jkl} is a vector of classroom controls (grade level, class size, percent minority, percent free/reduced-price lunch, percent LEP, and percent special education). Once again: Equation (4) is run four times, and the IQA variables are entered as binary (scores of three or higher compared to those of two or lower) for each rubric and for the IQA combined score.

For the fifth research question, I examine the role of instructional quality as the mechanism by which tracking affects achievement. Baron and Kenny (1986) proposed three conditions which must hold for mediation to be present: 1) the independent variable (track level) and the dependent variable (student achievement) must have a relationship, 2) the independent variable (track level) and the mediator (instructional quality) must have a relationship, and 3) the mediator (instructional quality) and the dependent variable

(student achievement) must have a relationship, holding the independent variable (track level) constant. This is illustrated in *Figure 3*.

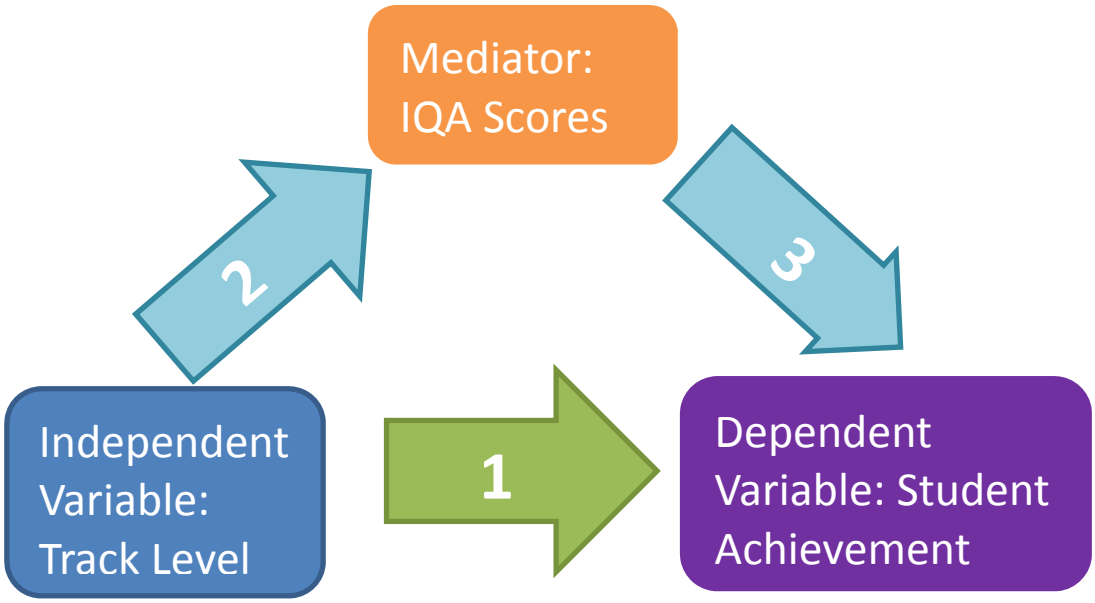

Figure 3:
Relationship between Track Level, Instructional Quality and Student Achievement

To estimate mediation using this approach I need a series of three models for each IQA variable. The first tests Baron and Kenny's first condition: the relationship between the independent variable (track level) and the dependent variable (student achievement). This relationship is tested in Equation (1), above. Because the IQA variables are not included in this step, there is only one model. The second set of models tests Baron and Kenny's second condition: the relationship between the independent variable (track level) and the mediator (instructional quality). This relationship is examined in the third

research question, shown in (3) above. The third set of models tests Baron and Kenny's final condition: a relationship between the mediator (instructional quality) and the dependent variable (student achievement), holding the independent variable (track level) constant.

While Baron and Kenny established the basic approach to testing for mediation in a linear regression approach, more recent research has extended this approach to working with multi-level data (Krull and MacKinnon, 1999, 2001). These researchers have pointed out that examining the impact of a group-level variable on an individual-level outcome will result in correlated error terms, violating a basic assumption of OLS. Therefore, following from the analysis in Research Question 4, these are four level models. Student achievement data is entered at the first level, and track-level variables and teacher IQA scores are entered at the second (observation) level. The participant and school levels include only participant-level and school-level intercepts. Student and classroom control variables are also included at levels one and two. My analysis is what Krull and MacKinnon (2001) call a "2-2-1" model: the intervention (track level) and mediator (instructional quality) are at the second level while the outcome (student achievement) is at the first level. This is shown in Equation (5):

(5)

$$\textbf{Level 1: } Achievement_{ijk} = \pi_{0jk} + \pi_{1jk}S_{ijk} + e_{ijk}$$
$$\textbf{Level 2: } \pi_{0jk} = \beta_{00k} + \beta_{01k}IQA_{jk} + \beta_{02k}Z_{jk} + \beta_{03k}Yr_{jk} + \beta_{03k}High\,Track_{jk} + e_{0jk}$$
$$\textbf{Level 3: } \beta_{00kl} = \gamma_{000l} + r_{00kl}$$
$$\textbf{Level 4: } \gamma_{000l} = \delta_{0000} + u_{000l}$$

This is an extension of the fourth research question, discussed above, with track level entered as a covariate. While the outcome of this particular set of equations is not

of direct interest, it is necessary to calculate the mediation effect, as I use the product-of-coefficients method (Zhang, Zypher and Preacher, 2008). This method multiplies the coefficient on the track level variable in Equation (3) by the second-level coefficient on the IQA variable in Equation (5), as shown in Equation (6):

$$\hat{\beta}_{1k} * \hat{\beta}_{01k} \tag{6}$$

This is done four times: once for each of the instructional quality variables. The standard error of this mediation effect is calculated as:

$$\sqrt{\left(\hat{\beta}_{1k}^2 * SE\left(\hat{\beta}_{01k}\right)^2\right) + \left(\hat{\beta}_{01k}^2 * SE\left(\hat{\beta}_{1k}\right)^2\right) + \left(SE\left(\hat{\beta}_{1k}\right)^2 * SE\left(\hat{\beta}_{01k}\right)^2\right)} \tag{7}$$

Because the coefficients in Equation (3) come from logistic regressions, while the coefficients in Equation (5) come from linear regressions, I must first scale the β_{1k} coefficients to make them comparable to the β_{01k} coefficients. To do this, I multiply each coefficient by the standard deviation of the high track variable and divide by the standard deviation of the appropriate IQA variable (Kenny, 2008; MacKinnon & Dwyer, 1993). If the coefficient in Equation (6) is found to be statistically significant, using the standard error from Equation (7), then there is a significant mediation effect of IQA on the relationship between track level and student achievement, indicating that instructional quality explains at least part of the relationship between track level and student achievement.

Descriptive Statistics on the Sample

The data from the Middle school mathematics in the Institutional Setting of Teaching (MIST) project includes student and teacher data from between 17 and 38 teachers in each district in each year. Because participating teachers left the study and

new participants were added, this amounted to 223 unique teachers and 9,847 students in the observed classrooms. In addition to the student, teacher and school demographic data, the main variables of interest for this analysis are the tracking variables, the IQA scores and student achievement. In this section I describe the sample in terms of each of these variables, and in the Results section, I examine how they relate to one another to answer my research questions.

Demographically, these districts were fairly typical of large urban districts in the United States. "Minority" students were often the majority, and 60 to 90% of students received free or reduced-price lunch. As shown in Table 1, district D had the highest percentage white students, with between 51% and 53% in each year. In each of these districts, about one-fourth to one-half of students were African American, but the majority of students in districts B and C were Hispanic (53% to 72%). District A was the only district with a significant minority (10 to 13%) of Asian and Native American students, captured in the "Other" column.

Table 1
Racial Makeup of Sample, by District by Year

Year	District	Black	Hispanic	White	Other	Total N
2007-2008	A	49%	21%	19%	10%	488
	B	30%	57%	12%	1%	588
	C	29%	69%	2%	0%	578
	D	40%	4%	53%	3%	707
2008-2009	A	44%	15%	30%	11%	579
	B	31%	61%	6%	1%	506
	C	27%	70%	3%	0%	636
	D	41%	4%	51%	4%	766
2009-2010	A	36%	16%	35%	13%	420
	B	36%	53%	10%	1%	799
	C	26%	71%	2%	1%	552
	D	39%	8%	48%	4%	690
2010-2011	A	32%	26%	31%	11%	553
	B	34%	57%	6%	2%	731
	C	26%	72%	2%	1%	489
	D	38%	6%	54%	3%	765

The percentage of test-takers classified as English Language Learners (ELL) varied greatly by district: from about 2% in district D to over 20% in district C. About 10% of students in each district received special education services. These districts also had significantly lower average achievement than the states in which they were located, by 0.5 to 0.7 standard deviations. This is particularly important to remember, as students' scores were z-scored to the state distribution, so the majority of students in this dataset have negative z-scores.

Each of these variables also differed substantially across schools, as shown in Table 2. While the average school was about 34% Hispanic, schools ranged from 0% to 99% Hispanic. Likewise, schools in the sample ranged from 0 to 100% Free/Reduced-Price Lunch, from 0 to 56% English Language Learners, and from 0 to 39% Special

Education. Although on average the sample was low-achieving, there were six schools with an average achievement above the state average, and one whose achievement was about three-quarters of a standard deviation above the state average.

Table 2
Range in School Averages across Variables of Interest

	Mean	Standard Deviation	Minimum	Maximum
Black	38%	19	0	80%
Hispanic	34%	31	0	99%
White	23%	23	0	73%
Other	5.4%	7.3	0	47%
Free/Reduced-Price Lunch	72%	23	0	100%
ELL	15%	13	0	56%
Special Education	11%	6.7	0	39%
Achievement z-score	-0.58	0.41	-1.43	0.76
N	118			

As shown in *Figure 4*, between 21% and 33% of students in videotaped classrooms were in untracked grade levels in each year. Fifty-three to sixty-one percent were in regular track classes, while only about 12% to 21% were in high track classes each year. This is about 300 to 550 students in high track classes across the four districts in each year. This is in part because the focus of the MIST project was on regular classroom instruction, so high track classrooms were under sampled. Sixth grade students were significantly less likely to be tracked ($p<0.01$) than 7th grade or 8th grade students (56.5% tracked in 6th grade as compared to 74.1% in 7th and 83.1% in 8th grade). Among tracked students, eighth graders were the most likely to be in high track classes (30.5%), due to the introduction of high school algebra offerings in eighth grade in many schools.

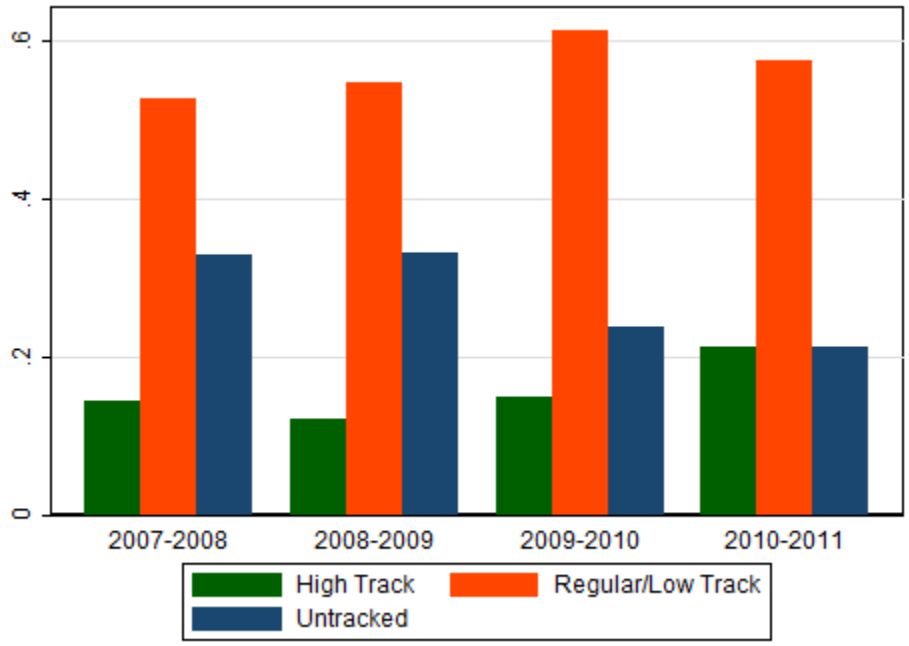

Figure 4:
Percent of Students in Each Track Level by Year

The average Potential of the Task score (2.91) was significantly higher than the average Implementation (2.37) or Discussion (1.76) scores ($p<0.01$). As shown in *Figure 5*, this is because the frequency of teachers with a Discussion or Implementation score of 1 or 2 is greater than the frequency with a 3 or 4, while the reverse is true for Potential of the Task scores. While the potential rigor of the tasks selected by most MIST teachers is focused on conceptual understanding (a score of three or higher), their actual implementation in the classroom is more likely to be a two or lower, focusing on procedural fluency alone. Likewise, the rigor of the concluding whole-class discussions is reduced to the sharing of answers, rather than discussion of strategies and pressing for generalizations. This aligns with Henningsen and Stein's (1997) findings that teachers reduce the cognitive demand of tasks when implementing them in the classroom, and

reinforces that examining all three rubrics is important to get a complete picture of rigor in the classroom.

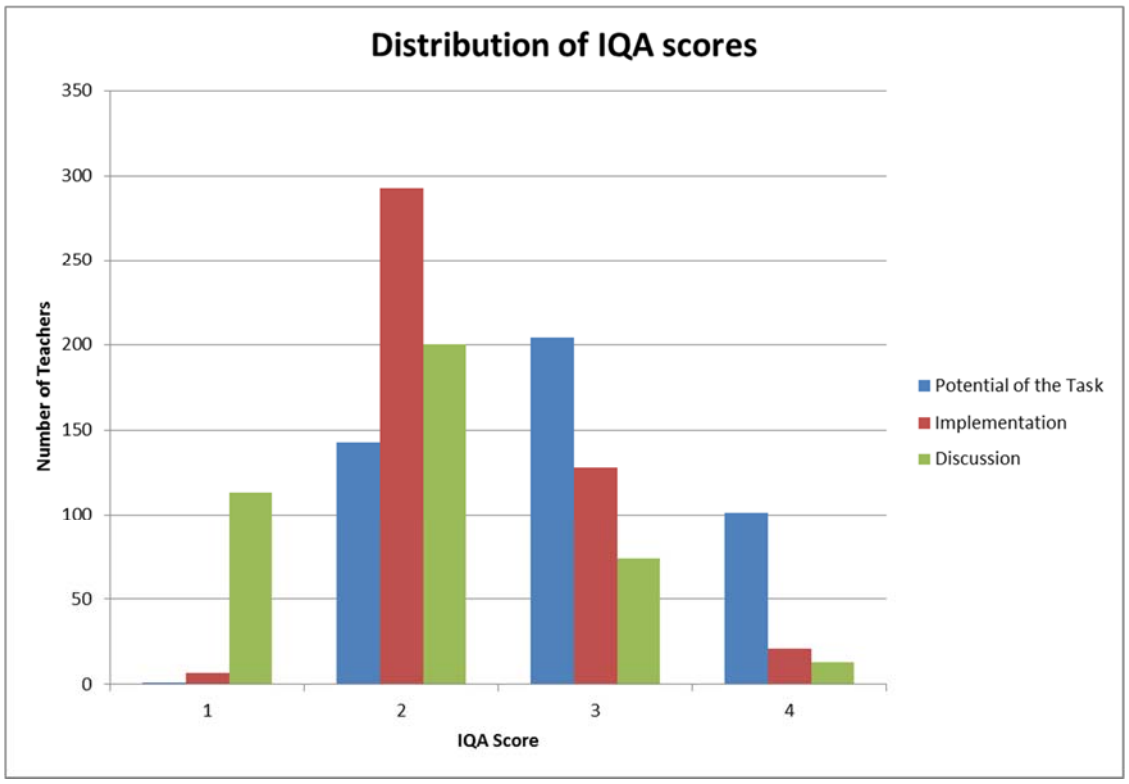

Figure 5:
Number of Teachers with Each IQA Score by Rubric across All Four Years

Teachers in the MIST study had a combined IQA score slightly above two, indicating that the average classroom is oriented toward learning and practicing rote procedures, rather than exploring and discussing multiple pathways to solving a complex problem. These average scores did not vary much across the four years of the study, though they did vary by district. In particular, District C had lower average scores on all rubrics than the other three districts. Districts A, B and D had similar average Potential of the Task scores (around a 3), indicating that teachers in those districts selected tasks

that required justification and explanation. However, District A had the highest implementation and discussion scores, averaging a 2.7 and 1.9, respectively. This indicates that more teachers in District A focused on explanation and justification through the concluding whole class discussion.

Results

The research questions addressed here will follow the paths illustrated in *Figure 6*. The first step is to establish whether there is a relationship between track level and student achievement, as drawn in path 1. Although found in previous research, corroborating this relationship in the MIST data is necessary for the justification of the research questions: if there is no relationship between tracking and student achievement, then there is nothing for instructional quality to mediate.

Second, I will address path 2 in my second and third research questions: are there measurable differences in instructional quality between teachers in tracked and untracked settings? Between high- and regular-track classrooms? Then, I will explore path 3 in addressing my fourth research question: Is rigorous mathematics instruction associated with higher student achievement? While the IQA has been linked to student achievement in other contexts (Resnick, Matsumura and Junker, 2006), other studies have found no relationship (Matsumura, 2008), and the relationship has not been established in MIST data. My final research question addresses the largest arrow (path 4): does instructional quality as measured by the IQA mediate the relationship between track level (high versus regular track) and student achievement?

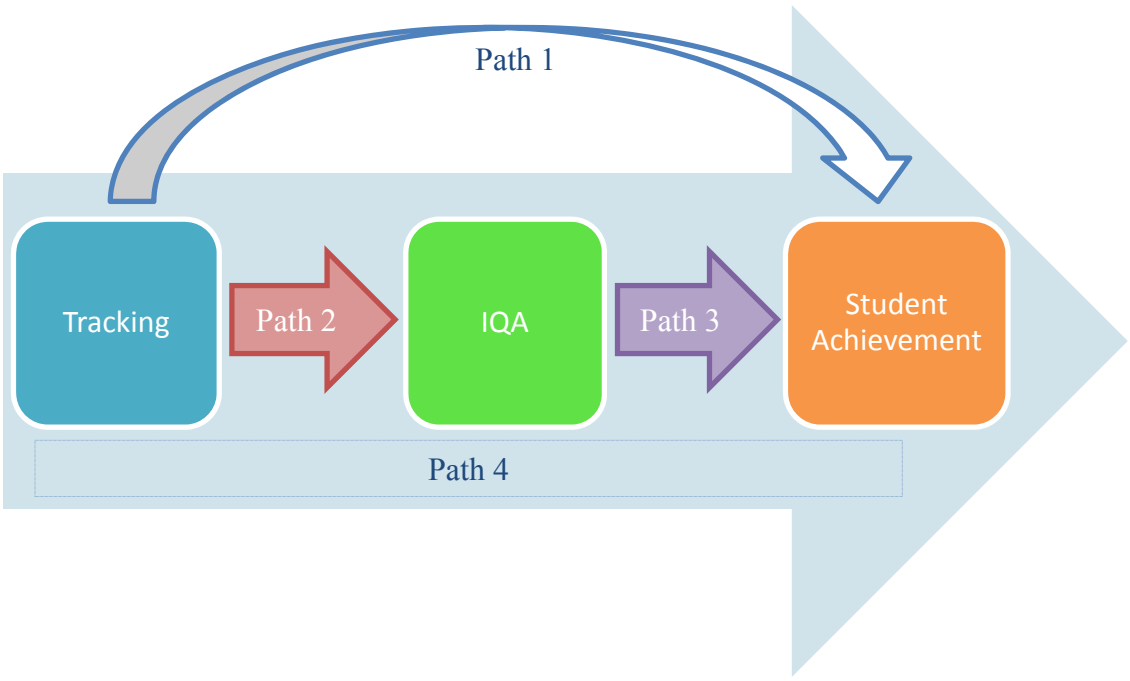

Figure 6:
Logic model of the analysis

Research Question 1: Are there measurable differences in student achievement by track level? As expected, there is a strong association between track level and student achievement in MIST schools. Illustrated in Figure 7, high track students had significantly higher achievement in all four years than regular track or untracked students, and untracked students out-scored regular track students during the same time. The scores of high track students also significantly declined over the four years of the study, while those of untracked and regular track students remained the same. During this same time, the total proportion of students in high track classes increased, which may explain some of the decline in scores: if high track classes were expanded to include lower-achieving students, the overall achievement of those classes would decline. This underscores the need for the year fixed effects in each of the models.

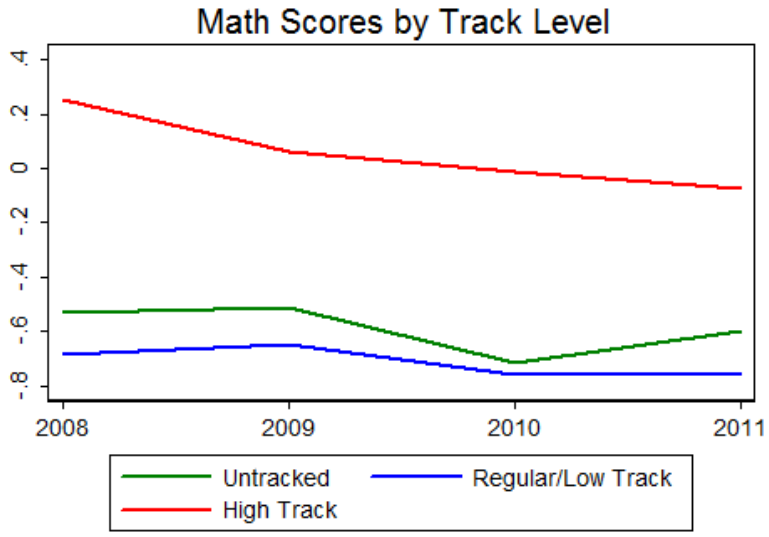

Figure 7:
Line Graphs of the Average Math Z-Scores in Each Year by Track Level

When examined using multi-level models with students nested within teachers, and teachers nested within schools, the size of the gap between high track and regular track students was about 0.8 standard deviations ($p < 0.001$). Controlling for prior achievement, race, free/reduced-price lunch, English Language Learner status (ELL), district and grade, high track students still out-performed untracked and regular-track students by about 0.12 standard deviations ($p<0.05$), but the difference between untracked and regular track students disappeared.

As indicated by the impact of introducing these variables to the model, race, socio-economic status, and language background are highly associated with both track level and student achievement. Black and Hispanic students are significantly less likely to be in high track classes, as are English Language Learners and students receiving Free or Reduced-Price Lunch. However, once prior achievement is controlled, race and FRPL status are no longer significant as predictors of track level. The strong inter-correlations

between achievement, track level and these demographic variables reinforce that they should be included in any models examining tracking and achievement, to allay the risk of omitted variable bias.

Research Question 2: Are there measurable differences in instructional quality between teachers in tracked and untracked settings? When examining a linear relationship between IQA scores and the presence of tracking, there is not a statistically significant relationship. There were few significant differences between tracked and untracked settings in average Potential, Implementation, Discussion or combined IQA scores, and those that were significant were often practically quite small, as shown in Figure 8. In general, unadjusted mean IQA scores were slightly lower in tracked settings than in untracked settings, but this difference was not statistically significant in all years.

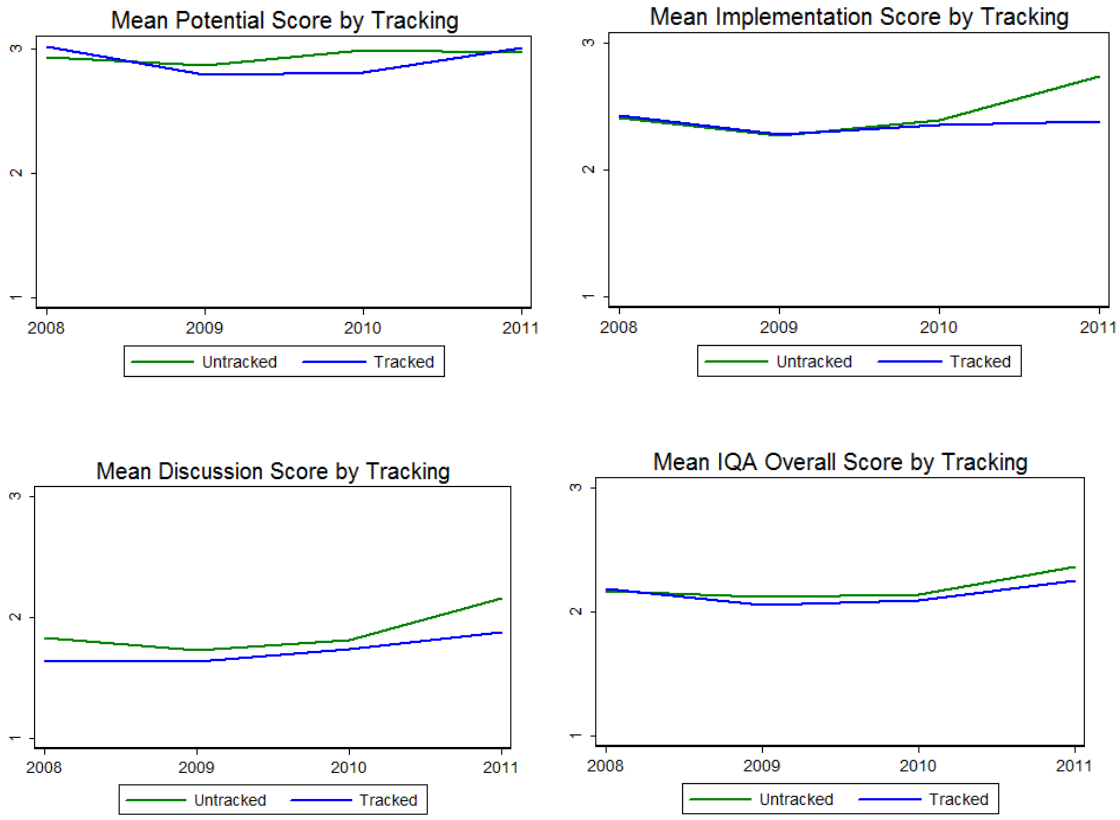

Figure 8:
Line Graphs of Unadjusted Average IQA Scores on Each Rubric in Tracked and Untracked Settings

Despite the lack of a linear relationship between mean IQA scores and tracking, there may be a difference between tracked and untracked settings in the propensity to select and implement high rigor tasks and discussions: those with scores of three or higher. This would occur if the distribution of scores were skewed negatively, with a few teachers drawing the average IQA score down significantly. *Figure 9* shows the proportion of teachers with a score of three or higher (high rigor) on each IQA rubric by tracked versus untracked classroom. This indicates that the proportion of teachers

selecting rigorous tasks is *lower* in tracked than in untracked classrooms. This is also true of discussion in all years except 2009, when there was no difference. The proportion of teachers with scores over three on implementation and on combined IQA is also lower in tracked classrooms in 2009 through 2011. The more consistent results here indicate that dichotomizing IQA scores into high and low rigor, in addition to being supported by the literature, is also supported by the data.

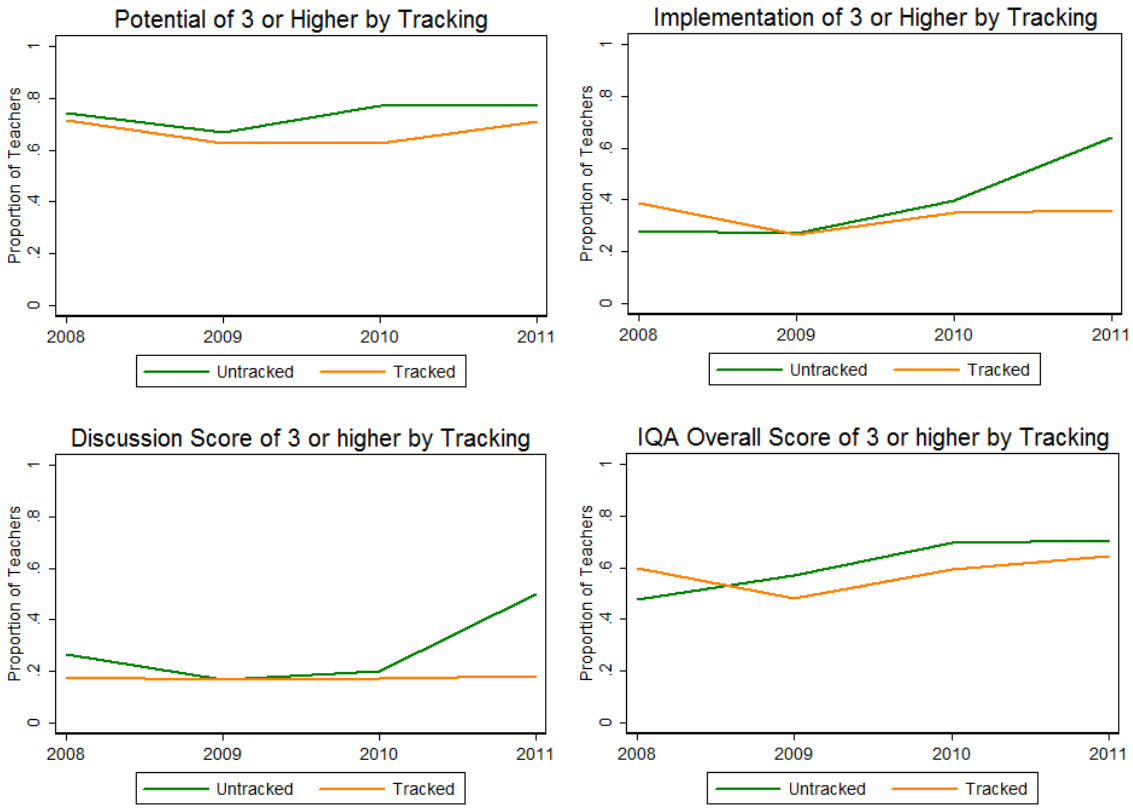

Figure 9:
Line Graphs of the Unadjusted Proportion of Teachers with High IQA Scores on Each Rubric in Tracked and Untracked Settings

The graphs in *Figure 9* show the proportion of teachers with high scores on the IQA rubrics unadjusted for district or classroom characteristics, however. To answer the second research question I dichotomized each IQA rubric into "high" and "low" (where high is a three or higher) before entering them into multi-level logistic regression models. The results of these models, in the form of odds ratios, are displayed in Table 3. This shows that, when district and classroom characteristics are controlled, the odds of a teacher implementing a rigorous task in tracked settings is nearly seven times *higher* than the odds of implementing a rigorous task in untracked settings ($p<0.001$). Likewise, the odds of a teacher having a high combined IQA score is about 3.8 times higher in tracked than in untracked settings ($p<0.001$). Thus, controlling for district and classroom characteristics reverses the relationship shown in *Figure 9*. The odds of a choosing a task with high potential or of carrying out a rigorous discussion was not significantly different between tracked and untracked settings. In all four models, the average propensity for high IQA scores varied significantly between participants, such that the odds ratios varied between participants by a factor between ten and sixteen.

Table 3:

Multi-level Logistical Regression Predicting the Odds of Rigorous Mathematics Instruction in Tracked and Untracked Settings

	Rigorous Task Potential		Rigorous Implementation		Rigorous Discussion		Rigorous Combined IQA Score	
Tracked	**0.70**	**(0.13)**	**6.76*****	**(1.44)**	**0.94**	**(0.23)**	**3.79*****	**(0.60)**
District 2	0.59	(1.20)	0.00**	(0.00)	0.03	(0.08)	1.71	(3.70)
District 3	0.00***	(0.00)	0.00***	(0.00)	0.00***	(0.00)	0.00**	(0.00)
District 4	2.59	(5.57)	0.00**	(0.00)	1.21	(2.97)	0.12	(0.27)
Year 2	0.35***	(0.04)	0.43***	(0.05)	0.19***	(0.03)	0.75*	(0.08)
Year 3	0.25***	(0.03)	1.97***	(0.24)	0.53***	(0.08)	1.45***	(0.16)
Year 4	0.22***	(0.03)	1.67***	(0.22)	0.67*	(0.11)	0.71**	(0.09)
Class is 7th grade	8.04***	(1.62)	6.65***	(1.82)	0.07***	(0.02)	1.45	(0.29)
Class is 8th grade	19.69***	(4.53)	3.56***	(1.01)	0.11***	(0.03)	1.36	(0.29)
% of class FRL	0.12***	(0.04)	0.10***	(0.03)	0.01***	(0.00)	0.05***	(0.01)
% of class LEP	289.94***	(157.2)	1.13	(0.70)	0.01***	(0.01)	17.85***	(9.66)
% of class SPED	0.11***	(0.05)	319.47***	(152.08)	0.00***	(0.00)	2.13	(0.91)
Class size	0.97**	(0.01)	0.96***	(0.01)	0.93***	(0.01)	0.96***	(0.01)
% of class minority	100.28***	(57.02)	0.44	(0.22)	68.65***	(40.44)	0.99	(0.49)
Random Effects:								
Participant Level	11.28***	(1.51)	12.82***	(1.83)	16.20***	(2.59)	10.11***	(1.42)
Observations	9847		9847		9847		9847	

Exponentiated coefficients (Odds Ratios); Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Research Question 3: Are there measurable differences in instructional quality between teachers in high- and regular-track classrooms? While teachers in tracked settings have higher odds of rigorous overall instructional quality than those in untracked settings, controlling for district and classroom characteristics, the analysis discussed in the previous research question combines the instruction across track levels. Here I will address whether, within tracked settings, instructional quality differs between high- and regular-track classrooms. Given the findings above, I will bypass modeling IQA scores as a linear relationship and treat the distribution of IQA scores as binomial, comparing high rigor (scores of three or higher) to low rigor (scores of two or lower). As Figure 10 shows, there is not a consistent relationship between track level and the unadjusted proportion of teachers selecting and implementing rigorous tasks or carrying out rigorous discussions.

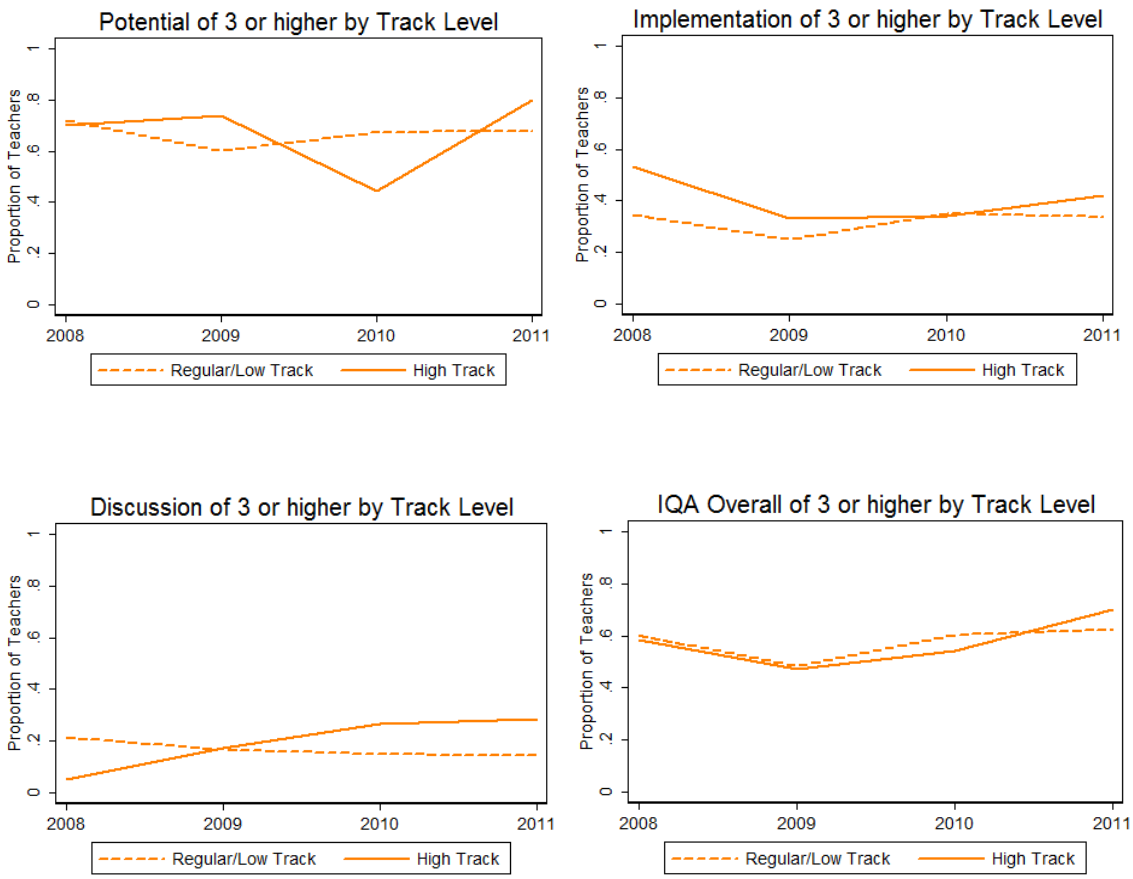

Figure 10:
Line Graphs of the Unadjusted Proportion of Teachers with High IQA Scores on Each Rubric by Track Level

However, Table 4 shows that, controlling for classroom characteristics, teachers in high track classes had significantly higher odds of choosing rigorous tasks (those with a Potential score of 3 or higher), implementing rigorous tasks (an Implementation score of 3 or higher), holding rigorous discussions, and having high combined IQA scores. Teachers in high track classes were 1.7 times more likely to select a task with high potential and more than 20 times more likely to implement a rigorous task when compared to teachers in regular track classes ($p<0.001$). The odds of a teacher holding a

rigorous discussion was also twice as high in a high track classroom as in a regular track class. Despite the appearance of the graphs above, when differences in classroom characteristics are controlled, students in high track classrooms are significantly more likely to be exposed to the type of mathematics instruction that is associated with high status knowledge than those in regular- and low-track classrooms. In this model, as in Research Question 2, there is significant variation at the participant level in the odds of scoring highly on the IQA, even controlling for track level and the other variables in the model.

Table 4:
Multi-level Logistic Regression Predicting the Odds of Rigorous Mathematics Instruction by Track Level

	High Potential of the Task Score		High Implementation Score		High Discussion Score		High Combined IQA Score	
High compared to Regular Track	**1.72****	**(0.31)**	**21.29*****	**(5.22)**	**2.07****	**(0.52)**	**11.93*****	**(2.62)**
District 2	0.18	(0.46)	0.00	(0.00)	0.03	(0.10)	0.75	(0.31)
District 3	0.00***	(0.00)	0.00***	(0.00)	0.00***	(0.00)	8.74***	(3.25)
District 4	1.02	(2.92)	0.01	(0.04)	1.37	(4.49)	0.01***	(0.00)
Year 2	0.31***	(0.05)	0.31***	(0.05)	0.11***	(0.03)	55.52***	(43.00)
Year 3	0.14***	(0.02)	1.58**	(0.25)	0.89	(0.19)	53.63***	(32.10)
Year 4	0.15***	(0.03)	1.02	(0.19)	0.69	(0.17)	0.86***	(0.01)
Class is 7th grade	34.19***	(10.46)	0.94	(0.31)	0.31*	(0.18)	63.01***	(41.42)
Class is 8th grade	126.47***	(42.33)	1.57	(0.54)	0.16***	(0.07)		
% of class FRL	0.01***	(0.01)	0.32**	(0.14)	0.02***	(0.01)	0.64	(1.49)
% of class LEP	828.24***	(633.35)	91.21***	(82.08)	0.00***	(0.00)	0.00***	(0.00)
% of class SPED	0.11***	(0.06)	158.44***	(326.44)	0.00***	(0.00)	0.15	(0.43)
Class size	0.96**	(0.01)	0.89***	(0.01)	0.81***	(0.02)	0.31***	(0.04)
% of class minority	107.79***	(75.72)	3.14	(2.15)	162.04***	(140.48)	0.50***	(0.07)
Random Effects								
Participant Level	9.95***	1.54	12.16***	(1.95)	18.70***	(3.31)	10.16**	(1.53)
Observations	7167		7167		7167		7167	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Research Question 4: Is rigorous mathematics instruction as measured by the IQA associated with higher student achievement? The relationship between IQA scores and student achievement (path 3 above) also appears to suffer from non-linearity. As shown in Figure 11, there is not always an obvious relationship in which increasing IQA scores are associated with increasing student achievement, unadjusted for prior achievement.

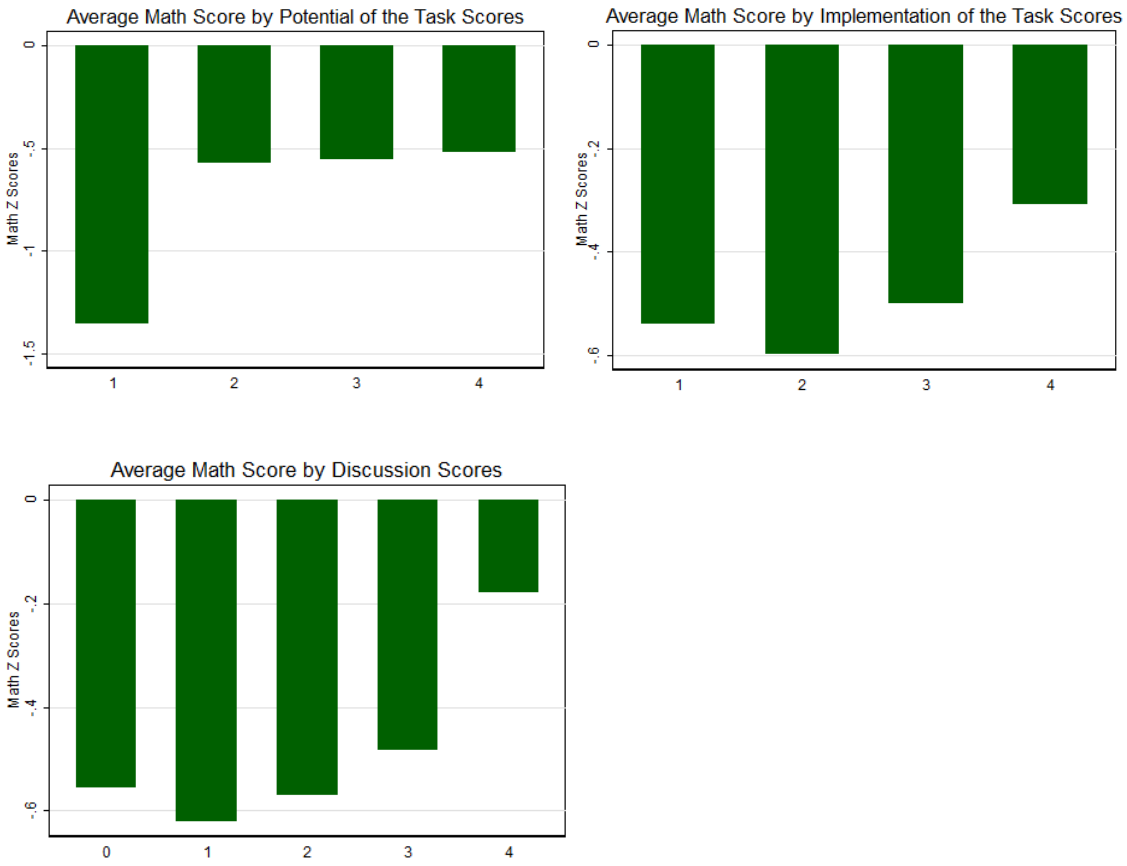

Figure 11:
Bar Graphs of the Unadjusted Relationship between Scores on Each IQA Rubric and Student Achievement

However, as discussed in the methods section, there is a theoretical justification for treating each IQA variable having a binomial distribution, with scores of two or lower denoted as "low rigor," while scores of three or four are "high rigor." When each IQA variable is introduced as a binary independent variable, there is a statistically significant, if small, relationship between high Potential of the Task and combined IQA scores and student achievement, as shown in Table 5. This relationship remains when prior achievement and student and classroom characteristics are controlled, and the size of the difference amounts to 5 to 6% of a standard deviation. This indicates that students whose teachers choose rigorous tasks are predicted to score about 0.05 standard deviations higher than similar students whose teachers do not select rigorous tasks. There is not a statistically significant relationship between high Implementation of the Task or Discussion scores and student achievement. In this model, there is significant variation in student achievement at all four levels, indicating that there are still unexplained differences between schools, participants, and classrooms, unaccounted-for by IQA scores or demographic variables.

Table 5:
Multi-level Regressions Predicting Student Achievement from Binary IQA scores

	Model 1 High Potential		Model 2 High Implementation		Model 3 High Discussion		Model 4 High Combined IQA	
High IQA Score	**0.05***	**(0.03)**	**0.02**	**(0.03)**	**0.04**	**(0.03)**	**0.06***	**(0.03)**
District 2	-0.09	(0.07)	-0.09	(0.07)	-0.09	(0.07)	-0.09	(0.07)
District 3	0.00	(0.07)	-0.00	(0.07)	-0.01	(0.07)	-0.00	(0.07)
District 4	-0.19**	(0.07)	-0.19**	(0.07)	-0.19**	(0.07)	-0.19**	(0.07)
Year 2	-0.01	(0.03)	-0.01	(0.03)	-0.01	(0.03)	-0.01	(0.03)
Year 3	-0.05	(0.03)	-0.05	(0.03)	-0.05	(0.03)	-0.05	(0.03)
Year 4								
LEP	-0.01	(0.03)	-0.01	(0.03)	-0.01	(0.03)	-0.01	(0.03)
FRL	-0.03	(0.02)	-0.03	(0.02)	-0.03	(0.02)	-0.03	(0.02)
Special Education	-0.20***	(0.03)	-0.20***	(0.03)	-0.20***	(0.03)	-0.20***	(0.03)
St's race minority	-0.10***	(0.02)	-0.10***	(0.02)	-0.10***	(0.02)	-0.10***	(0.02)
Prior Achievement	0.69***	(0.01)	0.69***	(0.01)	0.69***	(0.01)	0.69***	(0.01)
Class is 7th grade	0.06	(0.04)	0.06	(0.04)	0.06	(0.04)	0.06	(0.04)
Class is 8th grade	0.08*	(0.03)	0.08*	(0.03)	0.08*	(0.03)	0.08*	(0.03)
% of class FRL	-0.13	(0.08)	-0.12	(0.08)	-0.11	(0.08)	-0.12	(0.08)
% of class LEP	-0.14	(0.14)	-0.13	(0.14)	-0.13	(0.14)	-0.16	(0.14)
% of SPED	-0.09	(0.12)	-0.10	(0.12)	-0.08	(0.12)	-0.12	(0.12)
Class Size	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)	0.00	(0.00)
Constant	-0.03	(0.10)	-0.01	(0.10)	-0.02	(0.10)	-0.04	(0.10)
Random Effects								
School Level	-2.42***	(0.27)	-2.42***	(0.28)	-2.41***	(0.27)	-2.41***	(0.27)
Participant Level	-2.09***	(0.17)	-2.08***	(0.16)	-2.09***	(0.16)	-2.10***	(0.17)
Observation Level	-1.96***	(0.11)	-1.96***	(0.11)	-1.96***	(0.11)	-1.97***	(0.11)
Residual	-0.65***	(0.01)	-0.65***	(0.01)	-0.65***	(0.01)	-0.65***	(0.01)
Observations	6020		6020		6020		6020	

Standard errors in parentheses, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Research Question 5: Do these differences mediate the relationship between track level (high versus regular track) and student achievement? As described in the methods section, the first two steps of Baron and Kenny's approach to mediation analysis have already been established. First, a significant relationship between track level and student achievement, established in previous literature (e.g., Brewer, Rees & Argys, 1995; Gamoran, 1987), was supported in this data using multi-level modeling. Students in high track classes have predicted scores about 0.12 standard deviations higher, on average, than those in regular-track classes, controlling for student and classroom characteristics. Second, as shown in the third research question, there is a statistically significant relationship between track level and IQA scores, so that teachers in high track classrooms have higher odds of selecting and implementing rigorous tasks and holding rigorous discussions than those in regular-track classrooms.

The third step is establishing whether there is still a relationship between instructional quality and student achievement when track level is held constant. As shown in Table 6, the small relationship between Task Potential and student achievement remains statistically significant, even controlling for track level, but combined IQA scores are insignificant in this step[7]. Task Implementation and Discussion were both insignificant before track level was introduced as a covariate.

[7] These models include all the student-, classroom- and teacher-level covariates included in the models above, but their coefficients are excluded to save space.

Table 6:
Multi-level Regressions Predicting Student Achievement from IQA with Track Level as a Covariate

	Model 1	**Model 2**	**Model 3**	**Model 4**
Class is high track	0.12**	0.12**	0.12**	0.11**
	(0.04)	(0.05)	(0.05)	(0.05)
High Potential	**0.07***			
	(0.03)			
High Implementation		**-0.01**		
		(0.03)		
High Discussion			**0.01**	
			(0.04)	
High Combined IQA				**0.04**
				(0.03)
Observations	4307	4307	4307	4307

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Next, I estimated the size of the mediation effect of instructional quality using the product-of-coefficients method (Zhang, Zypher and Preacher, 2008). As shown in Table 7, the only mediation effect that was marginally statistically significant was the path of track level through task potential. About 8% of the total relationship between track level and student achievement can be accounted for by the greater likelihood of rigorous task potential in high track classes ($p<0.10$). Although more than 17% of the total effect was mediated by the combined IQA score, this mediation effect was not statistically significant because the standard error of the effect was large. These findings indicate that the type of mathematics instruction favored by the IQA mediates at best a small portion of the relationship between track level and student achievement.

Table 7:
Product-of-Coefficients method for Estimating IQA as a Mediator between Track Level and Student Achievement

	Total Effect of Track Level on Student Achievement		Mediation Effect through IQA		Percent of Total Effect Mediated
Task Potential	0.12**	(0.04)	0.009+	(0.005)	7.6%
Implementation	0.12**	(0.04)	-0.004	(0.019)	3.3%
Discussion	0.12**	(0.04)	0.001	(0.007)	1.2%
IQA Overall	0.12**	(0.04)	0.021	(0.016)	17.2%

Standard errors in parentheses
+p<0.10 * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Sensitivity Tests

Peer Effects. One potential rival explanation for the gap in student achievement and in instructional quality between high- and regular-track classrooms is peer effects. While there is similar achievement among students in the same track level, this analysis assumes that this similarity in achievement is not the sole reason for the relationship between tracking and student outcomes. If teachers adjust their instruction based on the average achievement of the classroom, then we can still argue that tracking is associated with instructional quality, and that instructional quality is a mediator between tracking and student achievement. However, if the average achievement of the class has an impact on individual achievement gains independent from its relationship with instructional quality, then it is important to separate this impact from the influence of instructional quality.

To test this, I added classroom average prior achievement to the models used to answer Research Questions 1, 3 and 4. As shown in Table 8, when included in the same model track level is no longer significantly associated with student achievement, but the

average achievement variable is significant ($p<0.01$). This suggests that there is no additional "labeling" effect of tracking: no effect of tracking over and above what comes from having a group of similarly-achieving students in a class together. This does not mean instruction does not matter, as teachers also know the prior achievement of students and may adjust their instruction accordingly. However, there *is* an additional effect of average achievement. The coefficient of 0.17 on this term indicates that two students in the same track level, but whose classes differ in their average prior achievement by one standard deviation are predicted to achieve about 0.17 standard deviations apart at the end of the school year. Simply put, regardless of the label of the class, the average achievement of the students in the class has a relationship with the students' predicted outcomes.

Table 8:
Multi-level Regression Predicting Student Achievement from Track Level with Classroom Average Prior Achievement as a Covariate

	Student Achievement	
	Coefficient	Standard Error
Class is high track	0.02	(0.05)
Classroom Avg Prior Achievement	0.17***	(0.04)
Observations	4307	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 9 shows the outcome of research question 3 when classroom average prior achievement is introduced to the models. Controlling for classroom average prior achievement, Task Potential, Implementation and Discussion scores are still significantly higher in high track classes. This indicates that high track teachers are significantly more

likely to select and implement cognitively demanding tasks and carry out rigorous discussions in high track classes than in regular track classes, even if the actual average prior achievement of the students is the same. So, peer effects have their own impact on teachers' instructional quality, but track level also has an independent affect.

Table 9:
Multi-level Logistic Regression Predicting IQA Scores from Track Level with Classroom Average Prior Achievement as a Covariate

	Potential of the Task	Implementation of the Task	Discussion	IQA Overall
Class is high track	1.54*	24.63***	7.09***	28.49***
	(0.32)	(6.82)	(2.19)	(7.34)
Classroom Average Prior Achievement	1.21	0.77	0.05***	0.24***
	(0.21)	(0.16)	(0.02)	(0.05)
Observations	7080	7080	7080	

Exponentiated coefficients; Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Finally, Table 10 shows how the outcome of Research Question 4 is affected by introducing peer effects. Controlling for classroom average prior achievement, Potential of the Task and IQA combined scores are still significantly associated with student achievement, and the size of the coefficients does not change substantially. This reveals that, although peer achievement is associated with an individual student's achievement, teachers' instructional quality also has a significant relationship, over and above the impact of peer effects.

Table 10:
Multi-level Regression Predicting Student Achievement from IQA Scores with Classroom Average Prior Achievement as a Covariate

	Student Achievement	Student Achievement
Potential of the task	0.06*	
	(0.03)	
IQA Overall		0.06*
		(0.03)
Classroom Average Prior Achievement	0.16***	0.15***
	(0.03)	(0.03)
Observations	6020	6020

Exponentiated coefficients; Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The impact of introducing peer effects to each of these models raises the question of whether classroom average prior achievement could be a more appropriate measure of tracking than the track level variable used here. Introducing this variable to question one shows that all of the impact of track level on student achievement can be accounted for by the fact that these classes have higher average achievement. On the other hand, the relationship between track level and rigorous instruction *cannot* be fully explained by the difference in average achievement in these classes, and neither can the relationship between rigorous instruction and student achievement. Therefore, while the average prior achievement of the class is associated both with achievement outcomes of individual students and some aspects of teachers' instruction, the track level variable here imparts more information than simply the impact of average achievement.

School and Teacher Fixed Effects. Another potential rival explanation for the effects found in this analysis is the non-random sorting of teachers and students. This can

happen at the school level, so that higher achieving students and higher quality teachers are more likely to be found in tracked than in untracked grade levels, or vice versa. This sorting can also happen at a classroom level, so that more qualified teachers and more able students are found in high track than in low track classes. In one sense, this is part of the main hypothesis of this paper: teachers with higher quality instruction will be sorted into higher track classes, and this non-random sorting is the mechanism by which tracking affects student achievement. On the other hand, it would strengthen the argument if these differences in gains could be shown to be due to the actual instructional quality, rather than other factors associated with the teachers or the schools they work in, such as their experience, education, school demographics or unobserved factors. The analysis in the main body of this paper attempts to account for the differences in students and classrooms by controlling for prior achievement and demographic characteristics, but it does not control for teacher or school characteristics except by partitioning the variance at the teacher level. Therefore, I tested each research question using school- and teacher-level fixed effects. School fixed effects analysis allows the comparison to be made between high- and regular-track classrooms in the same school. Teacher fixed effects analysis compares high- and regular-track classrooms within the same teacher. In other words, it looks at teachers that taught in high track classrooms in one year and in regular track classrooms in another year and compares their IQA scores and their students' achievement gains.

In the first research question, school- and teacher-level fixed effects both found nearly identical coefficients (0.13 and 0.11, respectively) as the main analysis, indicating that the gap in student achievement between high- and regular-track classes was about the

same within schools or teachers as it was overall. As shown in Table 11, in the second and third research questions, the coefficients were also nearly identical between the multi-level and the teacher fixed effects specifications. The relationship between track level and each IQA rubric was smaller in the fixed effects specification, indicating that some of the relationship between track level and teacher instructional quality may be due to pre-existing differences between the teachers. However, a larger portion of the relationship is due to changes in instructional quality within teachers. The coefficients were even smaller in the school fixed effects specification, signifying that a portion of the gap in instructional quality between settings is due to pre-existing differences between schools.

This does not change the conclusion that high track students are more likely to be exposed to rigorous instruction, even controlling for their own background and achievement and for the average achievement of the class. This simply indicates that the reasons teachers in high track classrooms have more rigorous instruction may be due in part to non-random sorting of teachers into these classes. Additionally, there is still a significant relationship in the fixed effects specification, indicating that some of the relationship is not due to permanent differences between teachers and schools, but to instructional changes when they are in high track classrooms.

Table 11:
Comparing Fixed Effects Logistic Regression to Multi-level Logistic Regression Predicting the Odds of Rigorous Mathematics Instruction by Tracking and Track Level

		Task Potential	Task Implementation	Discussion	IQA Combined
Tracked compared to Untracked	Multi-level model	0.70 (0.13)	6.76*** (1.44)	0.94 (0.23)	3.79*** (0.60)
	Teacher Fixed Effects	0.72 (0.13)	6.51*** (1.37)	0.94 (0.22)	3.64*** (0.57)
	School Fixed Effects	0.75** (0.03)	2.36*** (0.25)	0.47*** (0.05)	2.37*** (0.22)
High compared to Regular Track	Multi-level model	1.72** (0.31)	21.29*** (5.22)	2.07** (0.52)	11.93*** (2.62)
	Teacher Fixed Effects	1.68** (0.30)	19.12*** (4.58)	2.00** (0.50)	11.17*** (2.43)
	School Fixed Effects	0.88 (0.08)	1.43*** (0.13)	2.74*** (0.33)	1.88*** (0.17)

Exponentiated coefficients; Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In looking at the fourth research question using teacher fixed effects, I found that there was still a statistically significant relationship between the Potential of the Task selected by teachers and student achievement ($p<0.10$). This indicates that when teachers move from using low potential tasks to using high potential tasks, their students are

predicted to score about 0.05 standard deviations higher. However, this specification looks at the way teachers' instruction and their students' achievement change across years. While students' prior achievement and other demographics are included in the models, teacher fixed effects examine teachers across years and so across different groups of students. It is not possible to say whether the students' achievement improved because the teacher increased the rigor of her instruction, or if the teacher increased the rigor in response to having higher achieving students.

Using school fixed effects, Potential of the Task, Discussion and IQA Combined Score were all significantly related to student achievement. Within schools, students whose teachers had high Potential of the Task scores were predicted to out-score their peers whose teachers had low Potential scores by 0.04 standard deviations. The effect of Discussion was about the same size, while the coefficient on IQA Combined was about 0.05 standard deviations.

When I examined the final research question using teacher-level fixed effects, I found that the Potential of the Task score was still significant ($p<0.10$) as a mediator, mediating about 8% of the total relationship between track level and student achievement. Using school-level fixed effects, the IQA combined score was also significant as a mediator, mediating about 5.4% of the effect. The results of these sensitivity tests indicate that, within teachers, there is a gap between high- and regular-tracks both in their students' achievement and in the rigor of the instruction to which they expose those students. Some portion of the relationship between tracking and instructional quality is attributable to permanent characteristics of the teachers and schools, but most of it remains even when using teacher- and school-level fixed effects.

So, the instructional quality differences between high- and regular-track classrooms are due in part to the choice of teachers assigned to those classes, but largely to the instructional decisions those teachers make once they are there. Although only a small portion of the relationship between track level and student achievement is mediated through this definition of high quality instruction, this mediational effect is also found within teachers, so it cannot be attributed entirely to the assignment of more qualified teachers to higher tracks.

Limitations

An important limitation in this study is the small relationship between the measure of instructional quality (IQA) and student achievement. Although statistically significant, the size of the coefficients on Task Potential and the IQA combined score were practically quite small, and the coefficients on Implementation and Discussion were not statistically significant. This indicates that, although aligned with theory and qualitative research on what is high quality mathematics instruction, this measure is not well-aligned with the state tests used in these districts. This points to a much larger policy problem than can be addressed in this analysis: the lack of a relationship between what experts in the field identify as high quality instruction and what schools and teachers are being held accountable for. As the type of instruction valued by the IQA is associated with high-status knowledge, the gap in access to this type of knowledge may be associated with longer-term outcomes for students, such as college-going and professional careers. Future analyses should examine alternate outcomes such as these, as well as tests of reasoning that are better aligned with this definition of high quality instruction.

Nonetheless, in this era of high-stakes accountability, the gap between high- and regular-track students on these achievement tests begs for an explanation. Until the tests can be more appropriately aligned to high quality instruction, future analyses should investigate observational rubrics that are more closely associated with the achievement tests being used. For example, although this analysis examined the rigor of the tasks selected and implemented in the classroom, it did not assess the actual content of the tasks (i.e., *what* is taught, rather than *how* it is taught). It is possible for teachers to implement rigorous tasks that do not align with what will be asked on the end-of-year assessments. Likewise, the IQA focuses on the rigor of instruction in the classroom, but does not examine homework or other independent work. As the IQA rewards group work, and the state assessments require students to work independently, it is possible that the tasks a teacher assigns for students to work on independently may have as strong or stronger relationship with student achievement than those used in a group setting. In a 2008 study by Matsumura, Garnier, Slater and Boston, the assignments given by teachers in mathematics were found to be significantly associated with student achievement and with IQA scores. Future analyses should examine the role of task content and independent work in mediating the relationship between tracking and student achievement.

A second limitation is the lack of longitudinal data at the student level. The MIST project follows schools and teachers across time, but does not follow students. In this dataset, I have one prior year of achievement for each student, but I do not have access to data on what schools or classrooms the students were in prior to the year of analysis. Therefore, I cannot examine student-level fixed effects to see how students' achievement

and the instruction they are exposed to may change when they move from a regular- to a high-track class. This leaves the possibility of omitted variable bias in the non-random sorting of students. If students are more likely to be placed in high track classes because they more motivated or have more supportive families, then the differences in achievement attributed to track level and to instructional quality could be in fact due to background characteristics of the students. However, the third research question is less likely to be affected by this possibility. Teachers may adjust their instruction to be more rigorous if they perceive that their students are more motivated or more prepared, but as with the non-random sorting of teachers, this adjustment is part of the theory of the relationship between tracking and instructional quality. The reasons why teachers change their instruction may be because of the nominal level of the course, the actual prior achievement of the students, or their own perceptions of the students' motivation and background. Nonetheless, students in high track classes are being exposed to more rigorous instruction and more high status knowledge.

Conclusion

Previous research has established that students in high track classrooms significantly out-perform those in regular track classrooms, and many researchers have hypothesized that differences in teaching quality account for these gaps. While there has been a wealth of qualitative research documenting the differences in teaching practices and climate between high- and low-track classrooms (e.g., Oakes, 2005), there has been little research establishing how differences in instructional quality may mediate the relationship between track level and student achievement, particularly in mathematics. Starting from a definition of what counts as high quality mathematics instruction that is

based in both theory and these prior qualitative findings, the Instructional Quality Assessment provides an opportunity to quantitatively measure instructional quality and examine to what extent it serves as the mechanism by which grouping students for instruction is associated with differential student achievement.

This analysis found, as in previous studies, a significant relationship between track level and student achievement, even controlling for demographics and prior achievement. The size of this difference amounts to about 0.12 standard deviations, a medium-sized effect in educational research. This establishes that, in these four large, urban districts, there remains a significant gap in achievement gains between high- and regular-track students.

As a first step in examining whether instructional quality serves as a mediator in that relationship, I first confirmed that there are significant differences in instructional quality between track levels. I found that teachers in high track classrooms have significantly higher odds of choosing and implementing cognitively demanding tasks, holding rigorous discussions and having overall high-quality instruction than those in regular track classrooms. Teachers in high track classrooms were more than 20 times more likely to implement a rigorous task than teachers in regular track classrooms. This relationship remains when controlling for classroom average prior achievement, showing that high track teachers have significantly higher odds of rigorous instruction in high track classes than in regular track classes, even if the actual average prior achievement of the students is the same. In addition, I found that there was a small, but statistically significant relationship between instructional quality as measured by the IQA and student achievement. Students whose teachers had high Potential of the Task scores were

predicted to out-perform similar students whose teachers had low Potential of the Task scores by about 0.05 standard deviations.

Despite the statistical significance of the two parts of this relationship, only Potential of the Task was marginally significant as a mediator between track level and student achievement. About 8% of the total relationship between track level and student achievement could be accounted for by the greater likelihood of rigorous task potential in high track classes ($p<0.10$). None of the other IQA scores were statistically significant as mediators. The minimal mediation found here is likely due to the small size of the relationship between the measure of instructional quality and student achievement. Although there are large differences between track levels in the rigor of the tasks used and discussions held, these differences do not account for much of the gap in student achievement because the achievement tests are not highly correlated with these aspects of instructional quality. Future research should examine other facets of instructional quality, such as content taught and homework.

Nonetheless, the findings here do contain several important conclusions. First, gaps in achievement between track levels remain in these large, urban districts, even controlling for prior achievement and demographics. Second, students in high track classes are a great deal more likely to be exposed to tasks and discussions that require explanation and justification, rather than memorization or repetition. This supports the qualitative finding by Oakes (2005) that high track students are exposed to more high-status knowledge, thus preparing them for higher level courses and eventually professional careers where independent thinking is required. The correlation between race, socio-economic status and placement in these courses means that this rationing of

high-status knowledge for high-track students is likely to perpetuate existing inequalities. Therefore, even though the size of the mediation effect was small, the importance of the gaps in instructional quality between track levels looms large.

CHAPTER III

THE ROLE OF TEACHERS' VIEWS OF STUDENT ABILITY IN THE RELATIONSHIP BETWEEN TRACKING AND STUDENT ACHIEVEMENT

While Chapter II showed that students in high track classes have greater access to high-status knowledge, some opponents of tracking also argue that it is inequitable because it assumes that ability is a permanent state, internal to the student, and that achievement tests can accurately measure a student's innate ability (e.g., Abu El Haj & Rubin, 2009; Lotan, 2006). Even under more flexible modern tracking, students are usually assigned to course levels for at least a semester. This process assumes that students are "high" or "low" for at least this length of time, and that tests used to assess their level can reliably distinguish between high and low students. If this is the case, the rationing of high-status knowledge may be an efficient way to target instruction to the most deserving. In contrast, some researchers who argue against tracking use a "developmental" definition of ability that suggests that perceived ability is a relationship between the student's current position in a progression of capabilities and the opportunities afforded that student instructionally, rather than on their fixed position relative to other students (Oakes, Wells, Jones & Datnow, 1997; Watanabe, Nuñes, Mebane, Scalise & Claesgens, 2007). They contend that test results and grades are a reflection of performance at a given time, but not of a student's inherent ability (e.g., Abu el-Haj and Rubin, 2009; Horn, 2007; Zohar, Degani and Vaaknin, 2000). By sorting students into even semi-permanent groups, tracking forces students into a static category, while their actual ability is something mutable, that can and will change over time.

One major barrier to detracking efforts may be teachers' views about the nature of ability and the capabilities of their students (Abu el-Haj and Rubin, 2009). If teachers believe that ability is a permanent state and that test scores and grades are reliable measures of ability, then they may be unlikely to support detracking, and may assign students to within-class groupings that are as immutable as class-level tracks. Therefore, a developmental conception of ability is necessary to the success of detracking initiatives (Watanabe, et al., 2007). The implication of this argument is that detracking will not significantly affect achievement unless teachers in detracked settings conceive of ability as mutable rather than innate and fixed.

A great deal of qualitative research has addressed the issue of teacher beliefs about student ability, using largely case study and ethnographic methodologies (e.g., Hand, 2010; Lotan, 2006; Rist, 1970; Rubin, 2008). These studies have established that teachers' views of student ability differ between tracked and untracked settings, and that the way teachers think about student ability affects their interactions with students and thus affects student outcomes. However, this relationship has not been studied quantitatively: we do not know if these findings can be generalized to a larger population, and we do not know the size of the impact of teachers' views of student ability. Therefore, this study examines the relationship between teachers' views of student ability, tracking and student achievement, using a quantitative measure of the extent to which teachers hold a developmental view of student ability. This measure assesses the extent to which teachers have "productive" explanations for why students struggle or succeed in mathematics, productive views of student motivation, and describe productive supports for struggling students. Teachers with "productive" views discuss student

ability as mutable: they explain student struggle and motivation in terms of things that can be changed, rather than inherent or unchangeable factors. They also discuss supports that will allow students to develop a conceptual understanding of the material, rather than focusing only on developing procedural understanding. These productive views indicate that teachers believe their struggling students are capable of engaging in rigorous mathematical work.

"Productive" views align with the developmental conception of ability that detracking proponents argue is necessary to its success (Watanabe, et al., 2007), so using this rubric, I first examine whether this developmental conception of ability is found more prevalently in untracked or in tracked settings: 1) Do teachers in untracked settings have more productive explanations of why students struggle in mathematics than those in tracked[8] settings? 2) Do they have more productive views of student motivation? 3) Do they describe using more productive supports for struggling students? Following this, I examine the relationship between a developmental conception of ability and student achievement: 4) Are more productive views of students' mathematical capabilities associated with higher student achievement? Finally, I examine how the relationship between teachers' views and tracking may impact student achievement: 5) To what extent do teachers' views of students' mathematical capabilities moderate the relationship between tracking and student achievement?

[8] Because each teacher is interviewed only once, but they may teach both high- and low-track classes, I cannot distinguish between teachers' views of high- and low-track students.

Literature Review

Defining "Ability." In 1908, Alfred Binet developed an instrument designed to identify students in need of special education services and to measure their current capabilities, as a starting point for providing those services. This instrument was measured on an age-equivalent scale, where the age at each level was designed to be the youngest age at which "a child of normal intelligence should be able to complete the task successfully" (Gould, 1981: p. 179). Dividing this score by chronological age created the Intelligence Quotient, or IQ, that has formed the popular understanding of intelligence ever since. This scale has been largely misused. While Binet intended it to reflect something developmental and changeable, it has been used to measure innate and "natural" intelligence. This dichotomy drives discussion of ability to this day. On the one hand, ability can refer to something innate in people, inherited and uni-dimensional. From this perspective, ability does not change over time, it cannot be influenced by instruction or other outside factors, and it can be easily and reliably measured, though whether a test has been developed that can measure innate ability or not is up for debate (Gould, 1981; Oakes et al., 1997).

On the other hand, a developmental conception of ability sets aside the possibility of natural intelligence. Whether there is some form of intelligence that is inherited or not, "ability" is something that is mutable, influenced by instruction and other external factors, and therefore is not consistently and accurately reflected in achievement or IQ tests (Oakes, et al., 1997). While IQ and achievement tests often purport to measure a conception of ability that is fixed and innate, in fact there are serious shortcomings to this interpretation. From the beginning IQ tests have shown differences on the basis of race

and socio-economic status, indicating that they are not correlated only with innate intelligence, but may be vulnerable to racial and economic bias (Oakes, et al., 1997). Measuring innate intelligence was also not the original intention of IQ tests. Binet saw his instrument as an ongoing inventory of students' capabilities; something that could change over time and with adequate instruction (Gould, 1981). The reliance on achievement and IQ tests to separate students for instruction carries with it both the problems of the inaccuracy and possible bias of the tests and the focus on something innate in students, rather than a continuum of ability that can be influenced by instruction.

The Role of Teachers' Views of Ability in Tracking and Student Achievement. Both proponents and opponents of tracking often make the assumption that ability is fixed and can be accurately and reliably assessed. Proponents argue that it makes sense to separate students by innate ability and target resources to those who can make the most use of them (e.g., Hallinan, 1994). Opponents often argue that the gaps in the quality of education received are unfair to lower-ability students, or that standardized tests cannot reliably differentiate between the two groups, while also assuming that ability is fixed (e.g., Gamoran and Weinstein, 1998; Rosenbaum, 1976). In contrast to these arguments, some detracking proponents have contended that effective untracked education requires "significant shifts in *how we understand ability*" (p. 440: Abu el-Haj and Rubin, 2009, emphasis added), and teachers have not been supported to develop these shifts in belief and practice (Henningsen & Stein, 1997; Lotan, 2006; Loveless, 1999). Policymakers have argued that this poses a major barrier to detracking efforts. If, after a detracking initiative, teachers continue to believe that there are inherent ability differences between their students, and that "low ability" students cannot benefit from the

same cognitively demanding work that they give their "high ability" students, they are likely to continue to sort students on a classroom level and support their lower achieving students by lowering the cognitive demand of the tasks they give them. Therefore, a developmental conception of ability is necessary to the success of detracking initiatives (Abu el-Haj and Rubin, 2009; Lotan, 2006; Loveless, 1999).

Horn (2007) discussed the "ways in which conceptions of students, subject, and teaching are embedded in teachers' daily work" (p. 38). Horn described what she called the "Mismatch Problem," in which teachers believe that students with low prior achievement are not prepared for the rigorous mathematics of the type described in Chapter II, and so they feel torn between teaching the way they think they need to and the way the curriculum demands. Using ethnographic interviewing and observation in two high schools in California, Horn (2007) found that some teachers challenged the idea of "fast" and "slow" students. They viewed students' capabilities as flexible, and saw all students as able to engage in rigorous work when supported instructionally. However, other teachers talked about students as fast or slow, lazy or motivated, and saw students' ability to engage in the curriculum as directly resulting from those innate differences. These teachers felt that "slower" students required more repetition and could not handle work that required higher order thinking skills, such as justification and explanation. Because the school had eliminated remedial courses, teachers felt torn between teaching the way they felt they needed to and the way the curriculum demanded.

Zohar, Degani and Vaaknin (2000) also argue that teachers' beliefs about the type of material their students can handle have a strong influence on their practice, and that those beliefs can be a self-fulfilling prophesy: if teachers do not believe low-achieving

students can handle higher order thinking, they will not assign them tasks requiring it, which will make it less likely that those students will be capable of handling that type of work in the future.

This conflict between teachers' views of ability and the material they are expected to teach may be particularly problematic in untracked settings, where low- and high-achieving students are grouped together in the same classrooms. In a review of prior case studies and ethnographies across core subject areas, Abu el-Haj and Rubin (2009) argued that most teachers support detracking "in principle" but have difficulties implementing it in practice. The authors found that middle school teachers in detracked settings continued to think of students in terms of low, middle and high ability groupings within their classes. These teachers complained about the lack of resources to help "low ability" students who were not classified as special education and often attributed students' difficulty with the material to the culture of the neighborhood or family.

In discussions of the importance of teacher views of ability in detracking reforms, it is implicit that student learning will be affected by teachers' views. While a large body of research has examined the relationship between tracking and student achievement, and generally found no difference in achievement between tracked and untracked settings (e.g., Kulik & Kulik, 1982; Slavin, 1990), much less research has looked at the relationship between teacher beliefs about ability and student achievement. Those studies that do relate teacher views of ability and student achievement discuss it in terms of the "vicious cycle" and self-fulfilling prophesy of teacher beliefs, wherein teachers believe that some students cannot achieve at a high level, so they do not give them challenging work, and the students do not excel.

A seminal work in this area was Rist's (1970) study, "Student Social Class and Teacher Expectations: The Self-Fulfilling Prophecy in Ghetto Education," which followed the teacher and students in one kindergarten classroom and the development of ability groups in this classroom. He found that the kindergarten teacher assigned students to groups based on non-academic sources of information such as behavioral checklists, experience with older siblings, and her interactions with them in the first eight days of school. These groups strongly reproduced class boundaries between the students and were correlated with non-academic factors such as appearance and cleanliness. Once these groups were created, the teacher proceeded to direct the majority of her attention to the "Table 1" students (those she placed closest to the front), who she believed were "fast learners." Table 2 and 3 students received little verbal interaction with her and were often ridiculed or isolated from other students. These students responded by disengaging from instruction and/or acting out verbally and physically. There was no movement either into or out of the highest group once it was established, even after passing through first and into second grade. Although the first and second grade teachers used academic indicators to assign reading groups, these academic indicators were themselves a reflection of the unequal expectations and instruction the students received in prior grades. Thus, teacher expectations of students became a self-reinforcing cycle.

Subsequent researchers have found similar results to Rist. For example, Hand (2010) used participant observation to examine opposition and learning in one 8th grade mathematics classroom. Hand's work extended Rist's "Self-Fulfilling Prophesy" into the middle grades and examined how student behaviors and teacher expectations were reinforcing. She found that when students seemed to struggle, the teacher watered down

the mathematics, making it less about reasoning and more about recall. This lack of opportunity to engage deeply with the subject matter led to more confusion among the students, creating opportunities for opposition, even with very able students. Although both the Rist (1970) and Hand (2010) studies indicate that teachers' views of students can interact with students' own behavior and learning and become a vicious cycle of failure, neither used a specific measure of teachers' views to assess the size of their impact on student learning or differentiate between different aspects of teachers' views about student ability.

Aspects of Mathematics Teachers' Views of Student Ability. In a forthcoming study, Jackson, Gibbons, Garrison and Munter (2013) use the prior literature on mathematics education and student ability to explicitly outline three aspects of mathematics teachers' views that may matter for student outcomes. While the Jackson et al. (2013) study does not connect the teachers' views to tracking policies, it does discuss teachers' views about the nature of ability in ways that I will continue to use throughout this paper. The three aspects of teachers' views are: 1) explanations for why students struggle, 2) views of student motivation, and 3) views of supports for struggling students. In this study, Jackson et al. (2013) frame each aspect of teacher views of students' capabilities as existing along a continuum from "unproductive" to "productive." In general, productive views "position the teacher as able to effect change in his/her teaching that will support students, particularly struggling students, to substantially participate in rigorous mathematical activity," (Jackson, personal communication, 2012), while an unproductive view suggests students' capabilities are static, or influenced by factors outside of the teacher's control. Therefore "productive" views are those more

aligned with the developmental view of ability hypothesized to be necessary for the success of detracking initiatives (Watanabe, et al., 2007).

Teacher explanations for why students struggle. The first aspect of teachers' views of student ability is the explanations teachers provide for why students struggle in mathematics classrooms. On the one hand, teachers may argue that students struggle for reasons that are inherent to the student and outside the teacher's control, such as differences in students' inherent intelligence (an "unproductive" view). On the other hand, teachers may see students' performance in mathematics as arising from the opportunities they have been afforded. In this case, teachers see all students as capable of succeeding in rigorous mathematics with the right supports (a "productive" view). This reflects a developmental conception of ability. As such, I hypothesize that "productive" views of the explanations for why students struggle will support the success of untracked classrooms.

In discussing the relationship between teacher views and the success of detracking efforts, several studies explicitly or implicitly mention teachers' explanations for why students struggle. In her 1991 book, *Lower Track Classrooms*, Page examined curriculum, instruction and culture in lower track classrooms in two schools in one district. Page argued that prior research has focused on whether tracking "works" and occasionally on the process by which tracking might affect achievement, but has found few significant differences because it has not looked closely at curriculum and culture within schools and classrooms. Using participant observation in English and social studies classes, she found that teachers characterized low-track students as less skilled,

and their explanations of these students' difficulties emphasized out-of-school factors, such as poor home lives.

Twenty years later, Worthy (2010) interviewed twenty-five sixth grade English/Language Arts teachers who taught at least one "regular" and one "honors" class. She concentrated on how teachers "talked about students and instruction in the different class levels" (p. 27). She found that, although no questions were asked about ability or levels of classes, the majority of teachers talked about it anyway. Only four of the twenty five teachers held productive views, saying that instruction did not have to be lower for students in lower track classes. The remainder of the teachers talked about students in "regular" classes as having worse "work habits, behavior, ability and interest in learning" (p. 279), and they talked about these characteristics as "static and immutable" (p. 279). These teachers primarily blamed students' struggle in the classroom on their assumptions about students' home lives (most had not visited students' homes) and prior schooling experiences. In both cases, these were causes outside of the teacher's control, and therefore fall into what Jackson et al. (2013) label "unproductive."

Another set of studies on teachers' explanations for why students struggle focused on the context of detracked or untracked schools[9] or classrooms. In 1997, Oakes, et al., interviewed teachers and parents in ten schools undergoing detracking. The authors found that teachers in detracked schools often believed that intelligence is innate, fixed and uni-dimensional. These teachers saw "ability" differences as a legitimate reason to

[9] I will continue to use the term "untracked" in this chapter to refer to settings in which students are not grouped by ability, so as not to distinguish between schools that have never been tracked and those that were tracked and then "de-tracked." I will use the term "detracked" only when it was used by the original researchers to refer to a change in tracking policy.

separate students, and "ability" as something determined outside of school. They sometimes attributed the struggles students faced to "culture" or race (euphemistically called "demographics"). In the detracked schools in this study, teachers had often been exposed to the idea that ability and intelligence are developmental, and tried to assimilate these views into their practice, but they often failed to do so. They talked about "multiple intelligences" or "learning styles" as explanations for differences between students in a superficial way, but they still ranked and classified students. These insufficiently productive views of students' capabilities (as more fixed than flexible) prevented true detracking.

The Page (1991), Worthy (2010) and Oakes et al. (1997) studies show qualitatively that teachers in tracked and untracked settings may both espouse unproductive explanations for why students struggle. Page (1991) and Worthy (2010) make the implicit argument that these views are tied to teaching in a tracked setting, but Oakes, et al. (1997) showed that teachers in untracked settings can continue to hold unproductive views. None of these studies used a specific rubric for measuring teachers' views or examined how these views were related to student achievement outcomes. Thus, further research is needed to compare teachers' views in tracked and untracked settings, as well as how those views may interact with tracking to affect achievement.

Teacher Views of Student Motivation. The second category of teachers' views of student ability stems from one popular explanation for why students struggle: student motivation. In Jackson et al.'s (2013) study, teachers in "unproductive" schools tended to talk about motivation in the same ways they talked about ability: as determined by factors outside of their control, such as inherent laziness or a dislike of mathematics. However,

teachers in "productive" schools saw motivation as something that the quality of their instruction could influence.

In examining tracking and teachers' beliefs, a few studies have examined the relationship between tracking and how teachers discuss motivation. In 1984, Finley used her own observations as well as interviews with her former colleagues to examine how students and teachers were allocated to different course levels in the English department (Finley, 1984). She observed that teachers spoke more positively about student motivation in high-track classrooms, and teachers of high-track classes tended to characterize motivation as intrinsic to students.

More than 20 years later, Reed (2008) conducted interviews and classroom observations with two National Board Certified mathematics teachers, both teaching two different tracks of high level mathematics. Reed found that students in regular calculus and pre-calculus were seen as less motivated than those in honors. Additionally, teachers saw more behavioral problems in the regular classes, and attributed these behavioral problems to factors outside their teaching, whereas in higher track classes teachers attributed behavioral problems to students' need for more challenging material.

In 2005, Carbonaro argued that prior research showing that high-track students have higher achievement gains has ignored student effort. He defined effort as "the amount of time and energy that students expend in meeting the formal academic requirements established by their teacher and/or school." (p. 28). The author found that being in a higher track was associated with higher levels of 10th grade effort, as reported by the teacher. Although the author argued that his findings demonstrated that the higher gains in high-track classrooms may be influenced by student effort, the use of teacher

reports is key. Another interpretation of Carbonaro's findings is that teachers perceive their high-track students as putting forth more effort. Without data from the students or an outside source, it is unclear whether these students actually do try harder, but it is clear that the researcher (and possibly the teachers) saw this effort as outside of the teacher's control.

In short, research on how teachers talk about motivation has shown that many teachers in tracked settings hold unproductive views: attributing greater inherent motivation to those in their high track than their low track classes. However, there has been little research on the views of teachers in untracked settings, or on how views of motivation may interact with tracking to affect achievement.

Teacher views about supports for struggling students. The final category of teachers' views of student ability is supports teachers advocate using with struggling students. Jackson et al. (2013) characterize the type of supports teachers describe providing to struggling students as "productive" and "unproductive," wherein productive supports are those that attempt to include struggling students in instruction focused on developing a conceptual understanding of mathematics, while unproductive supports focus only on developing procedural understanding. Productive supports, by including struggling students in rigorous mathematics with their peers, rather than separating them for remedial instruction, acknowledge that struggle is not a permanent state, caused by inherent low ability. Instead, all students struggle at some point, and the role of the teacher is to support students to continue to engage in rigorous mathematics with their classmates. Unproductive supports, by contrast, assume that some students, due to their inherent low ability, cannot succeed in rigorous mathematics with their "high ability"

classmates, and so instruction must focus on developing procedural fluency alone rather than conceptual understanding. Because "productive" supports promote including all students in rigorous mathematics, I hypothesize that these supports are also necessary to the success of untracked classrooms.

Several studies have examined the types of supports teachers provide to students who struggle, both in tracked and untracked settings, but few have examined how teachers talk about these supports. One body of research has compared the instruction in high- and low-track classrooms, generally using small case-study methods. Worthy (2010) found that lower track teachers "freely talked about lowered expectations and watered down instruction" (p. 279). Likewise, Reed (2008) found that teachers said they used less group work in regular than in honors classes because they saw these students as behavior problems during group work time. Although Reed (2008) and Worthy (2010) argue that their studies indicate a lower level of instruction in low-track classes, they did not examine how teachers adjust instruction within, rather than between classes.

Other studies have discussed how teachers deal with the variety of incoming achievement levels within their untracked classrooms. Lotan (2006) maintained that teachers have to plan for heterogeneity when planning lessons, not adjust the lessons once in the classroom. Lotan found that teachers often apply only one measure of competence (achievement on tests and grades), and that is picked up on by the students, who apply it to themselves and their peers. Similarly, Rubin (2003) found that grouping practices used by teachers to ability group students in untracked high school classrooms were sometimes obvious to students and made for uncomfortable situations.

Finally, Rubin (2008) used case studies of three schools to examine teachers' views of students and their instructional practices in detracked settings. The author only discussed how teachers dealt with variation in student achievement within their classrooms in one school. In Elmtown (a suburban, racially and socioeconomically mixed school), teachers saw detracking as part of an effort to build community and provide more opportunities to students. There was a wide range of achievement levels in the school, and prior to detracking high- and low-tracks were highly segregated by race. Detracking was seen as a way to help low-achieving students and to achieve more classroom equity. Teachers tried to connect the material to students' lives and make it engaging; they tried to create assignments that were flexible and so could be used at different levels by all. They addressed class and race as valid topics in a social studies classroom and often employed class discussion.

In examining the supports teachers provide to struggling students, most prior research has focused on examining the differences in practices between high- and low-track classrooms, an issue addressed in Chapter II. While this establishes that teachers feel the need to teach differently to students they see as struggling (low-track students), it does not establish how teachers behave when low- and high-achieving students are grouped together in untracked classrooms, nor does it examine how teachers talk about the supports they provide and why they provide them. As discussed above, some researchers argue that productive supports of struggling students are those that include them in rigorous mathematics along with their classmates (Jackson et al., 2013). Further research is needed on whether teachers in untracked settings espouse such supports, and

whether an endorsement of those supports interacts with heterogeneous grouping to affect student achievement.

Thus, although recent discussions of detracking have begun to focus on how teachers discuss the nature of "ability" and its role in sorting students for instruction, more research is needed on the link between teachers' views, tracking and student achievement. While many conceive of ability as something innate and fixed, some researchers argue that ability develops over time and can be influenced by instruction. Although the small case study research discussed here has described the importance of teachers' beliefs about ability, these findings have not used a measure of teachers' views that is aligned with a developmental conception of ability to quantify the impact of teachers' beliefs on the relationship between tracking and student achievement.

Additionally, this "developmental" conception of ability is viewed by some as necessary to detracking efforts because it allows teachers to see their instruction as influencing student success and motivation and to extend supports to struggling students that allow them to engage in rigorous mathematics (e.g., Abu el-Haj and Rubin, 2009; Lotan, 2006). This implies that the lack of a significant difference in student achievement between tracked and untracked settings found in much of the tracking research may stem from inattention to the role of teacher beliefs in moderating the effect: detracked students will not have higher achievement unless their teachers also hold developmental views of ability. The case studies discussed here have not established whether this is the case.

Research Questions

Using the MIST sample of 224 middle grades math teachers from 30 schools in four large urban districts and a rubric discussed below, this study will address the following research questions: 1) Do teachers in untracked settings tend to have more productive explanations of why students struggle in mathematics than those in tracked settings? 2) Do they have more productive views of student motivation? 3) Do they describe using more productive supports for struggling students? While policymakers argue that developmental views of ability are necessary to the success of detracking reform, prior research indicates that teachers in untracked settings may not be more likely to hold these views than teachers in tracked settings. These research questions will use a rubric designed to differentiate between developmental and fixed views of student ability to establish whether teachers in untracked settings are more or less likely to hold developmental views.

My fourth research question directly addresses the issue of whether teacher beliefs are associated with student achievement: 4) Are more productive views of students' mathematical capabilities associated with higher student achievement gains? While Rist (1970), Hand (2010) and others have shown a relationship between the way teachers talk about their students and those students' achievement, this analysis will measure the consistency and the size of that relationship in a large multi-state sample.

Finally, several authors (e.g., Horn, 2007) argue that a developmental conception of ability is necessary to the success of detracking efforts. This argument implies that untracked classrooms will not have higher achievement than tracked classrooms unless the teachers in those classes also hold this developmental view, and that teacher views

may "matter more" in untracked than in tracked settings. Thus, my fifth research question is: 5) To what extent do teachers' views of students' mathematical capabilities moderate the relationship between tracking and student achievement? As shown in Table 12, I hypothesize that teachers with productive views who teach in untracked settings will have significantly higher student achievement than those with unproductive views or those in tracked settings. I do not have a hypothesis for whether unproductive views in untracked settings or productive views in tracked settings will be associated with higher achievement, but I expect that the highest achievement will be found in untracked classrooms where the teachers have productive views, and the lowest achievement will be found in tracked classrooms where the teachers have unproductive views. Therefore, I hypothesize that productive views of student ability are a necessary condition for the "success" of untracked settings, but they may not be a sufficient condition to make tracked settings successful.

Table 12:
Hypothesized Interaction between Tracking and Teachers' Views

		Tracking	
		Untracked	*Tracked*
Teachers' views of student ability	*Productive*	Highest Achievement Outcomes	Middle
	Unproductive	Middle	Lowest Achievement Outcomes

Data and Measures

The data for the analysis in this chapter also come from the Middle school mathematics and the Institutional Setting of Teaching (MIST) study at Vanderbilt University. Because teachers could be interviewed who were not observed, there were 224 unique teachers who were interviewed in one or more years. One hundred eight teachers were interviewed in only one year, 38 were interviewed in two years, 35 were interviewed in three years and 43 teachers were interviewed in all four years of the study. These teachers were interviewed in person in January of each year by a graduate student or principal investigator on the MIST study. The focus of these interviews was on the teachers' vision of high quality mathematics instruction, the supports they provide to students and how the district's theory of action for improving mathematics instruction is playing out in their schools. The data from these interviews was used in scoring teachers on the VSMC rubrics.

Views of Student Mathematical Capabilities. To measure teachers' views about student ability, I use the Views of Student Mathematical Capabilities (VSMC) rubric. Using the aspects of teachers' views discussed above (Jackson et al., 2013), Jackson developed the VSMC rubrics to rate teachers' views of students in three areas: 1) Nature of explanations of student performance, 2) Views of student motivation, and 3) Nature of instructional supports. Throughout this analysis I will refer to scores on these rubrics as Explanations, Motivation and Supports. These rubrics are grounded in research on how teachers categorize students (e.g., Horn, 2007) and will serve to measure the size of relationship between teachers' views, tracking policy, and student outcomes.

Teachers' views in each category are scored as Productive, Unproductive, or Mixed. "Productive" views "position the teacher as able to effect change in his/her teaching that will support students, particularly struggling students, to substantially participate in rigorous mathematical activity," (Jackson, personal communication, 2011) while an Unproductive view sees students' capabilities as static, or as influenced by factors outside of the teacher's control. Therefore Productive views are those more aligned with the developmental view of ability hypothesized to be necessary for the success of detracking initiatives (Watanabe, et al., 2007). Mixed views include teachers who alternate between Productive and Unproductive views of student capabilities when talking about different groups of students and those who may be transitioning from an Unproductive into a more Productive view or vice versa.

The VSMC rubrics were applied to interview transcripts from all four years in the summer of 2011, with an update to the coding process resulting in significant re-coding during the summer of 2012. While several interview questions were particularly fruitful for VSMC coding, rubrics were not applied only to certain questions on the interview protocol, but rather to key concepts that may have been discussed. Therefore, coders were provided with a list of both potential keywords (e.g., motivation, adjust, different, and challenge) and of interview questions where they might find codeable material. Table 13 shows a few examples of interview questions.

Table 13:
Sample Questions from the MIST Teacher Interview used to Rate Views of Students' Mathematical Capabilities (VSMC)

1. What are the most important challenges of teaching mathematics in your school?
2. In your classroom, when students do not learn as expected, what do you find are typically the reasons?
3. Are all of the students in your classes motivated? *If not*, why not?
4. Do you feel you need to adjust your instruction for different groups of students within a class? Why or why not? *a. If so,* for which groups of students? How do you adjust your instruction?
5. Is there anything that you would like to do instructionally that you feel you can't do in your classroom? a. Why do you feel you can't _____ in your classroom?

Raters scanned the responses to each of the potential interview and follow-up questions and then searched the transcript for key words. When they encountered a key concept, coders applied the VSMC rubrics to the instance, rating a "turn of talk" (a teacher's response to a question) as Productive, Unproductive or Mixed. For example, a male teacher in district A in the first year of the study was asked "When your students do not learn as expected, what do you find are typically the reasons?" In response, he said:

> I think part of it is that they don't believe they can learn. They have, they have come to the conclusion that they aren't capable of learning and therefore they are unwilling to put up the effort… it goes back to what I was saying at the very beginning about how the problem is this academic status, getting the kids truly intellectually engaged in what's happening. So that's why the nature of the task is so important. How can tasks be designed that can draw everybody in and get their input?

This statement was rated "Productive" on the *Explanations* rubric because the teacher positions the students' struggle and capability as something he can influence through his own instruction. It is not rated on the *Supports* rubric because, while the teacher mentions the nature of the tasks, he does not describe what he does to design tasks that can "draw everybody in."

On the other hand, a female teacher in district B in the fourth year of the study was asked about how she adjusts her instruction for students who are struggling, and she responded:

> I try and keep it as, as simple, but because it's really hard to give a really hard, difficult, higher thinking lesson to the class as a whole. And so what'll I'll usually do is make it doable because I don't want them to get discouraged. I don't want them to think this is above me. I can't do this. I don't want them to just give up and so I usually give them something a little bit more simple.

This was rated as "Unproductive" on the supports rubric because the teacher says that she simplifies the task for students, avoiding the "higher thinking lessons" and thereby reducing the cognitive demand of the tasks for students she saw as less capable.

In general, segments were not coded on more than one rubric: if a particular statement by a teacher was coded for *Explanations* of student struggle, it was not also coded for *Supports* for struggling students. Likewise, while each of the questions in Table 13 were on the interview protocol in at least one year, they were not all present across all four years. Because the nature of teachers' beliefs was not an explicit focus in all years, interviewers occasionally skipped these questions and often did not probe teachers deeply on their views, resulting in responses that could not be coded for VSMC. For example, if a teacher were asked about the most important challenges in his school and he replied "Basic skills and prior knowledge," this could not be rated on the VSMC

"Explanations" rubric because it is not clear what he sees as the source of that problem. However, if the interviewer followed up by asking why that was a problem with his students, he might explain it in terms of the instructional opportunities they have been provided (Productive) or in terms of external factors outside of his control (Unproductive).

Teachers were rated as "Mixed" if they 1) had some Productive and some Unproductive statements, or 2) alternated between Productive and Unproductive within one statement. This included teachers who talked about most students as capable of rigorous work, but made exceptions for some students; teachers who said that all students were capable most of the time, but made exceptions for some situations; and teachers who explicitly said that all students could be supported to participate in rigorous instruction in response to one question and explicitly referred to low ability students as less capable of cognitively demanding work in response to another question. This combination of reasons for having a Mixed score makes the interpretation of the score difficult, as will be discussed below.

On the Supports rubric, supports for students receiving special education services and those for English Language Learners (ELLs) were treated differently from other supports. Because special education supports are often mandated under Individualized Educational Plans (IEPs), these supports were not rated as Productive or Unproductive, as they were not expected to stem from the teacher's own beliefs about what was necessary for supporting struggling students. Supports for English Language Learners were rated on a separate "Supports for ELLs" rubric with the same categories of Productive, Unproductive and Mixed. This was done because there is a separate body of research on

the types of supports that are productive for ELL students in mathematics. As it is not the focus of this analysis, I do not use this rubric.

Four coders coded all interview transcripts in 2011, with two as "expert" and two as "novice" coders. At the beginning of the summer, the novice coders received a full day of training on the rubrics and practiced coding transcripts with the expert coders present. Then, the novice coders were assigned a set of transcripts to code, all of which were double coded by an expert. The novice and expert met to discuss each transcript and any discrepancies in coding and to deepen the novice coder's understanding of the rubrics through practice. Finally, novice coders were assigned transcripts to code independently. Every one to two weeks, 10% of the novice's interview transcripts were selected and double coded by an expert. Throughout coding, reliability was maintained at 80% overall and 70% on each code. In 2012 this process was repeated, separating the Supports for ELLs from the general Supports rubric. Each transcript that had been coded for Supports was re-coded to check if these supports referred to English Language Learners. If so, they were re-coded on the Supports for ELLs rubric.

Tracking and Student Achievement. In addition to the in-person interview, each district also provided MIST with the student achievement data, class enrollment files and demographic data for the students of each of these teachers. This analysis will make use of tracking variables and student achievement from these files. As in the analysis in Chapter II, although individual classes could be "high" or "regular" track, the existence of one "tracked" class per grade level indicated that all students in that grade were separated by ability. In other words, if some sixth grade students in a school were

in honors classes, by default all other sixth grade students in that school were ability grouped by virtue of not being mixed in with those honors students.

The student achievement outcome used in this analysis is students' scores on the state assessment, z-scored to the state distribution. Student demographics, such as race, gender, ELL, special education and free/reduced-price lunch status will be used as covariates. School-level demographics, such as the racial composition, percent free/reduced-price lunch and the percent meeting or exceeding expectations on the state test were obtained from publicly-available data on district websites. Finally, teachers responded to a survey each year that included their demographic information such as years of experience, race and number of math courses taken.

Methods

In examining the relationship between teachers' beliefs and tracking in the first three research questions, I must account for other characteristics of classrooms, schools and districts that may also be correlated with tracking and teachers' views. Prior research has established a relationship between tracking and classroom, school and district characteristics: tracked classes and schools tend to be larger and more racially and socio-economically diverse, and higher grades are more likely to be tracked (Lucas and Berends, 2002). If these characteristics are also correlated with teachers' views of students' mathematical capabilities (VSMC), excluding them from the models would result in omitted variable bias. Therefore, I employ grade and district fixed effects and control for the percent of students receiving free or reduced-price lunch at the school and classroom levels, percent white at the school and classroom levels and school and class size. Additionally, I examine school fixed effects as a sensitivity test. These models do

not include teacher controls in these research questions because I do not want to account for the reasons why teachers may vary in their VSMC. The question is if students in untracked settings are more or less likely to be exposed to teachers with Productive views, regardless of the reasons why their teachers may hold these views.

To address the first three research questions, the dependent variables are the VSMC scores for Explanations, Motivation and Supports. I use multinomial logit models with "Productive" as the baseline category. The structural model for this analysis is of the form:

(8)

$$y_i^* = \beta_1 Tracked + \beta_2 C + \beta_3 S + \beta_4 G + \beta_5 D + \varepsilon$$

Here y_i* represents the latent, or underlying, propensity for a teacher to be rated as "Unproductive" or "Mixed" as compared to "Productive" on the VSMC rubric *i.* *Tracked* is a binary variable indicating whether the grade level is tracked or untracked, *C* is a vector of classroom characteristics (the percent free or reduced-price lunch, percent white and class size), *S* is a vector of school characteristics (the percent free or reduced-price lunch, percent white and school size), and *G* and *D* are grade and district fixed effects. Using multinomial logistic regression, I estimate the difference in probability of having Productive explanations of students' performance, views of student motivation, and supports for struggling students between teachers in tracked and untracked settings.

In addressing the fourth and fifth research question, I must account for teacher characteristics in addition to the school and district characteristics discussed above. While teacher characteristics such as age, math degree and certification have not been shown to be correlated with student achievement (Giglio, 2010), others such as years of experience and number of math courses taken have been shown in some cases to have an

impact (Kukla-Acevedo, 2009). Likewise, some of these same qualifications have been shown to be correlated with teacher beliefs (e.g., Brady & Woolfson, 2008) and with tracking (e.g., Heubert and Hauser, 1999), such that teachers with a math background may be more likely to see rigorous mathematics as necessary for all students (Brown and Gray, 1992; Horn, 2007) and more qualified teachers may be more likely to teach in tracked settings (Brookings Institution, 2009). If these teacher characteristics are associated with both the VSMC measure and student achievement, excluding them would lead to omitted variable bias. In other words, I could attribute differences in student achievement to differences in teachers' beliefs, when they are in fact due to differences in teacher experience or education or other unobservable differences. Therefore, the safest approach is to employ teacher fixed effects, examining the relationship between tracking, beliefs and student achievement within teachers. In the MIST data, teachers were interviewed in each of four years, and teacher fixed effects may be used where teachers switched from teaching in untracked to tracked settings across years, or vice versa. However, this approach severely reduces the sample, as there are few teachers who switched settings. Therefore the main analysis examines the full sample with controls for teacher race, experience and number of math courses taken, and a sensitivity analysis examines school and teacher fixed effects.

For the fifth research question, I have already examined the relationship between tracking and achievement in the MIST data in another analysis (Schmidt, 2013). This analysis showed that, as found in prior research, there was no statistically significant difference in student achievement between tracked and untracked settings. However, it is possible that this analysis masked a relationship that is moderated by teachers' views of

student ability. In other words, students in untracked settings may only outperform tracked students if they also have teachers with Productive views (those that endorse a developmental conception of ability). Therefore, I examine the interaction of teachers' VSMC scores with tracking and how that is associated with student achievement. As mentioned above, I am only able to employ teacher fixed effects on a smaller subsample of the MIST data, so I also examine a regression including teacher, student and school covariates, as well as grade and district fixed effects. This model also uses clustered standard errors at the classroom level, as students are expected to be more similar in achievement within classrooms. The model takes the following form:

(9)

$$Y_i = \beta_0 + \beta_1 Tracked + \beta_2 VSMC_{mixed} + \beta_3 VSMC_{productive} + \beta_4 Tracked * VSMC_{mixed} + \beta_5 Tracked * VSMC_{productive} + \beta_6 X + \beta_7 S + \beta_8 T + \beta_9 G + \beta_{10} D + e$$

Y_i is the achievement of student *i, X* is a vector of student control variables (race, free/reduced-price lunch, special education status, English Language Learner status) that have been shown to be correlated with student achievement and tracking (e.g., Gamoran, 2009; Oakes, 2005), *S* is a vector of school controls, *T* is a vector of teacher controls, and *G* and *D* are the grade and district fixed effects. The comparisons of interest are 1) the impact of having a teacher with Productive views in an *untracked* setting as compared to having a teacher with Productive views in a *tracked* setting, 2) the impact of having a teacher with *Productive* views in an untracked setting, as compared to having a teacher with *Unproductive* views in an untracked setting, and 3) the impact of having a teacher with *Productive* views in an tracked setting, as compared to having a teacher with *Unproductive* views in an tracked setting. The first comparison answers the question of whether untracked students can outperform tracked students when they have a teacher

with Productive views of students' mathematical capabilities. The second and third comparison answer whether having a teacher with Productive versus Unproductive views "matters more" in tracked or untracked settings. For the first comparison, I will use a linear combination of parameters to add the β_1 and β_5 coefficients. The second comparison will require only the β_3 coefficient, and the final comparison will combine the β_3 and β_5 coefficients. Therefore, the question of whether having a teacher with Productive views "matters more" in tracked or untracked settings is the first interaction term:

(10) $(\beta_3 + \beta_5) - \beta_3 = \beta_5$

Descriptive Statistics of the Sample

Table 14 shows some demographics of the teachers in this study across the four districts. Teachers in District A were older and more experienced than the teachers in any of the other districts. They also had taken significantly more methods and content courses than teachers in District B, and more advanced math courses than teachers in any of the other districts. They were also the most likely to be white and to have a full certification, rather than a partial or temporary certification.

Table 14:
Mean and Standard Deviation of Teacher Demographic Variables by District in all Four Years

	District A		**District B**		**District C**		**District D**	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Experience								
Yrs Taught Math	13.9	7.62	6.24	6.70	9.40	8.10	8.48	9.39
Total Yrs Taught	16.7	8.88	6.89	6.97	10.8	8.44	9.31	9.64
Courses Taken								
# of Methods	3.95	1.39	2.52	1.87	3.10	1.81	3.50	1.60
# of Math content	3.44	1.71	3.16	2.03	3.43	1.95	4.06	1.33
# Advanced Math	3.21	2.20	2.38	1.88	2.44	1.88	2.07	1.52
Age	46.3	11.4	36.5	9.63	42.1	8.99	36.9	13.3
White	91%		57%		22%		85%	
Full Certification	98%		88%		96%		82%	
N	444							

As shown in Table 15, in terms of student body, the MIST sample is typical of large, urban districts. In districts A, B and C the majority of students were non-white, with nearly 98 percent non-white in District C. The majority of students in all districts received Free or Reduced-Price Lunch, between 4 and 20 percent were Limited English Proficient and 8 to 10 percent received Special Education Services. The percent of students in tracked grade levels (those with more than one level of mathematics) varied greatly by district, with only about 3 percent of students tracked in District C, and more than half tracked by grade level in District A. The percent of students in high track classes was between 20 and 27% in each district. Since the students' achievement test scores were z-scored to the state distribution, the average score in each district was

between one third and one half of a standard deviation below the state average (-0.35 to -0.56).

Table 15:
Student Demographics of the MIST Sample across all Four Years

	District A	District B	District C	District D
African American	39%	32%	30%	40%
Hispanic	20%	55%	68%	5.3%
White	30%	11%	2.0%	51%
Other	12%	1.6%	0.57%	3.6%
Free/Reduced-Price Lunch	58%	65%	87%	77%
Limited English Proficient	20%	8.5%	20%	3.6%
Male	48%	51%	52%	53%
Special Education	9.8%	7.8%	8.8%	9.7%
Untracked	51%	2.9%	15%	40%
High Track	20%	27%	21%	26%
Regular / Low Track	29%	70%	64%	34%
Prior Year's Achievement	-0.27	-0.47	-0.36	-0.48
Current Year Achievement	-0.35	-0.55	-0.50	-0.56
N	9,698	8,989	10,888	12,493

Teacher demographics varied with the teachers' VSMC scores. As shown in *Figure 12*, the average years of experience is higher among teachers with Productive Explanations and Supports scores than among those with either Mixed or Unproductive scores. The pattern is the same for Motivation scores, but the differences are much smaller.

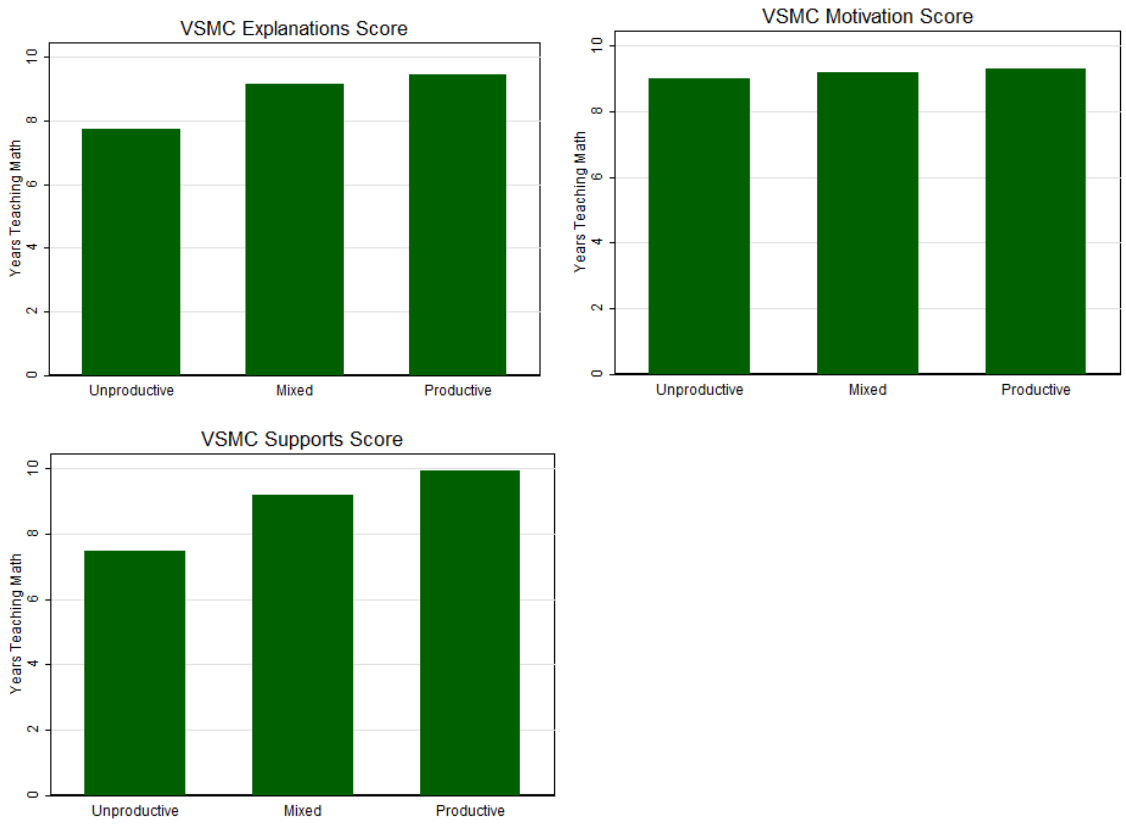

Figure 12:
Unadjusted Average Number of Years Teaching Math by the Teacher's VSMC Scores

There were not significant differences in the number of math methods or content courses taken by VSMC score, but the number of advanced math courses taken did vary by VSMC, as shown in the kernel density plots in *Figure 13*. In these plots, the green line represents teachers with Productive scores, the red line is teachers with Unproductive scores, and the blue dashed line is teachers with Mixed scores. The horizontal axis is the number of advanced math courses taken and the vertical axis is a probability density function based on the frequency of teachers at that level. So, the first graph shows that, between zero and two advanced math courses there are fewer teachers with Productive scores than teachers with Unproductive or Mixed scores (the green line is below the red

and blue lines). The reverse is true between four and six advanced math courses (the green line goes above the red and blue lines). There are also fewer teachers with Productive Motivation scores who took two or fewer advanced math courses, and more who took four or more courses. In Supports scores the relationship is less clear, with the green line for teachers with Productive scores below the red and blue lines at two to three courses, above both at more than four courses, and at about the same level at zero or one advance math courses. While there seems to be a stair-step relationship between years of experience and VSMC scores, in the number of advanced math courses taken, teachers with "Mixed" and "Unproductive" scores are similar, while teachers with "Productive" math scores have taken more courses.

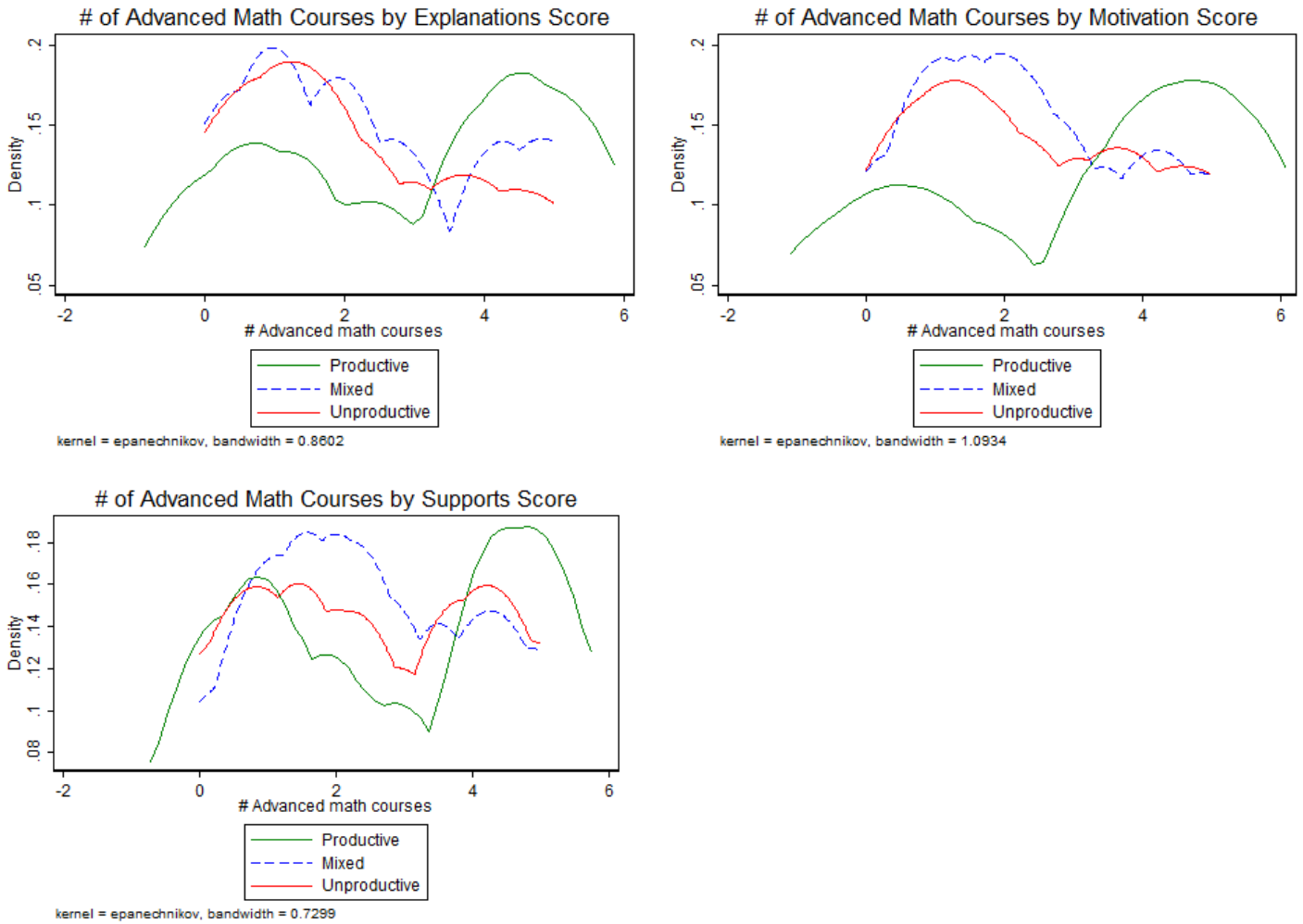

Figure 13:
Kernel Density Plots of the Number of Advanced Math Courses taken by VSMC Score

There was also a relationship between teacher race and VSMC scores, in which teachers with Productive Explanations or Motivation scores were significantly less likely to be white, and those with Unproductive Explanations or Motivation scores were more likely to be white ($p<0.001$). On the other hand, teachers with Mixed Supports scores were the most likely to be white, and there was no significant difference between Productive and Unproductive in the proportion of teachers who were white (see *Figure 14*).

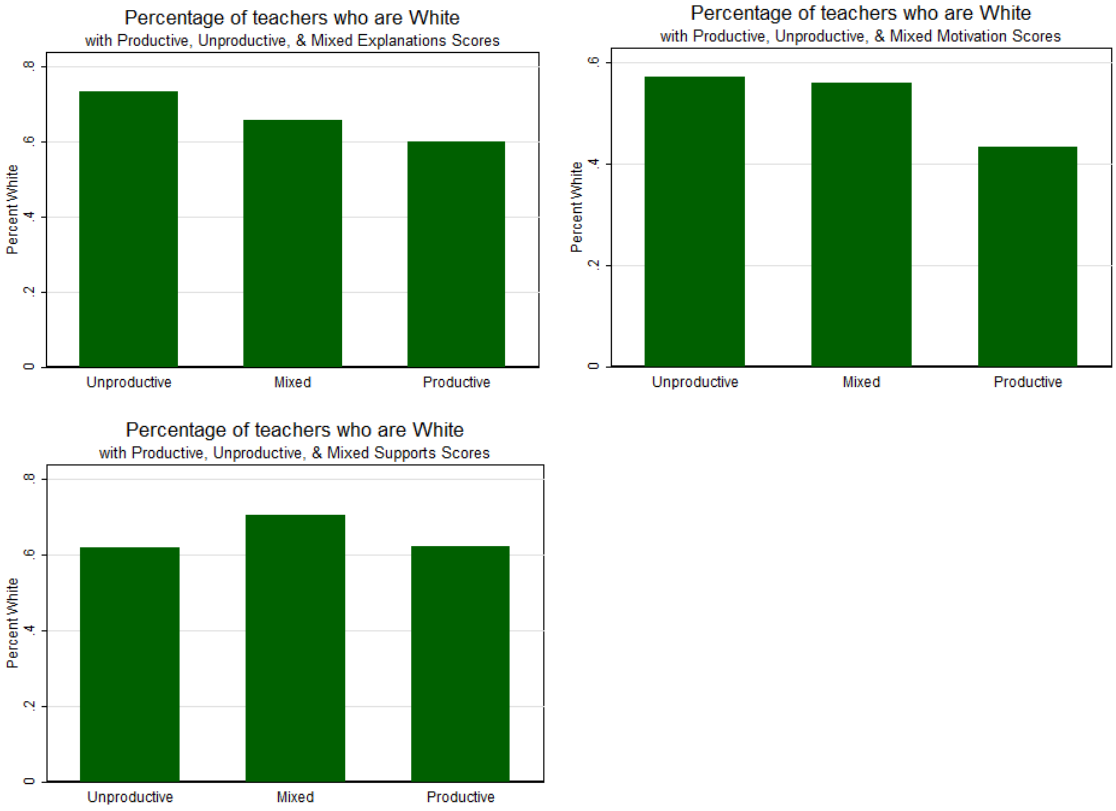

Figure 14:
Percent of Teachers who are White by VSMC Score

In addition to teacher-level variables, student demographics are correlated with tracking and VSMC scores. As shown in Table 16, teachers with Productive Explanations, Motivation and Supports scores had higher proportions FRL students and LEP students than teachers with Unproductive Explanations, Motivation and Supports scores. Teachers with Productive Explanations and Motivation scores also had higher concentrations of African American and Hispanic students and low proportions of white students in their classes. Teachers with Productive Supports scores also had higher proportions of Hispanic students, but lower proportion African American students in their classes. These differences in student composition between teachers with Productive and Unproductive VSMC scores were statistically significant at $p<0.01$. The difference between teachers with Mixed views and teachers in the other two categories varied in size and direction, and were not always statistically significant.

Table 16:
Student Demographics by their Teacher's VSMC Scores

		FRL	LEP	Black	Hispanic	White
Explanations	Unproductive	65.6%	9.5%	31.7%	31.8%	32.7%
	Mixed	71.3%	10.4%	38.1%	29.1%	28.3%
	Productive	70.2%	13.1%	34.3%	35.5%	24.7%
Motivation	Unproductive	67.6%	10.3%	31.7%	41.0%	23.7%
	Mixed	71.0%	13.3%	37.9%	36.4%	21.2%
	Productive	75.4%	14.4%	36.0%	44.7%	16.5%
Supports	Unproductive	68.5%	10.1%	39.3%	31.3%	25.7%
	Mixed	66.8%	11.2%	35.6%	26.8%	31.4%
	Productive	75.4%	18.6%	33.7%	43.6%	18.2%
Observations: 2,538						

Finally, some school characteristics were correlated with teachers' VSMC scores. Unsurprisingly, as VSMC was associated with student race, free/reduced-price lunch

(FRL) and achievement, these scores were also associated with the concentrations of FRL, minority and low-achieving students in the school. Teachers with Unproductive Explanations or Supports scores came from schools with significantly higher concentrations of white students, higher average achievement and lower concentrations of FRL students than teachers with Productive Explanations scores ($p<0.01$). Teachers with Unproductive Motivation scores also came from schools with significantly higher concentrations of white students and lower concentrations of FRL students, but the average school achievement was not significantly different.

Results

Research Questions 1 – 3: Tracking and Teachers' Views of Student Mathematical Capabilities.

As shown in Figure 15, before controlling for classroom and school characteristics, students in tracked settings were more likely to have teachers with Productive Explanations scores, but less likely to have teachers with Productive Motivation or Supports scores.

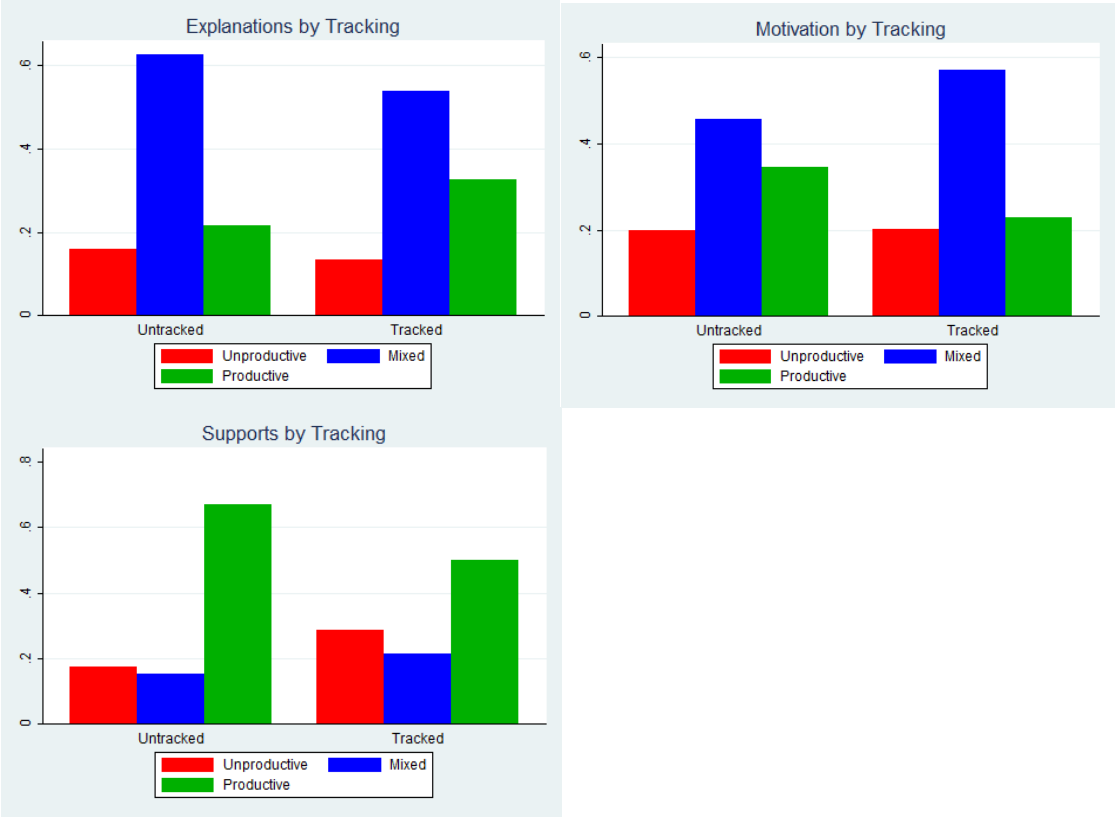

Figure 15:
Unadjusted Proportion of Teachers with Unproductive, Mixed and Productive VSMC Scores by Tracking

When controlling for classroom and school characteristics this remains the case. In Table 17, the comparison group is the "Productive" score on each rubric, so students in tracked settings had a significantly lower likelihood of having teachers with Unproductive explanations of why students struggle, but a higher likelihood of having teachers with Unproductive views of student motivation or teachers who described Unproductive supports for struggling students. Students in tracked settings also had a significantly lower likelihood of having teachers with Mixed Explanations of student

performance, but higher likelihood of having teachers with Mixed views of student Motivation and Supports.

Table 17:
Multinomial Logistic Regression Predicting VSMC Scores from Tracking

	Explanations	Motivation	Supports
Unproductive			
Tracked	-0.263**	0.734***	0.251***
	(0.087)	(0.115)	(0.076)
Constant	-1.016	-1.66***	6.593***
	(0.522)	(0.742)	(0.427)
Mixed			
Tracked	-0.229***	2.175***	0.558***
	(0.052)	(0.111)	(0.063)
Constant	-0.894**	4.529***	-3.053***
	(0.331)	(0.581)	(0.432)
Observations	15937	8618	16183

Standard errors in parentheses
* p < 0.05, ** p < 0.01, *** p < 0.001
These models control for the classroom, school and district variables described above

This is shown graphically in *Figure 16* using relative risk ratios. While the "risk ratio" would indicate the odds of having an Unproductive score as compared to a Productive score, for example, the "relative risk ratio" compares the risk ratios for tracked and untracked settings. The red bars indicate the increase in the relative risk ratio of Unproductive versus Productive scores for tracked as compared to untracked students. A ratio of less than one indicates that being in a tracked setting is associated with a lower risk of an Unproductive score, as compared to a Productive score, while the opposite is true for a ratio over one. The relative risk ratios for the "Productive" category are all one, because this is the comparison group. So, being in a tracked setting is associated with a

lower risk of having a teacher with Unproductive or Mixed Explanations for why students struggle or succeed in mathematics. On the other hand, students in tracked settings have about twice the odds of having teachers with Unproductive views of student motivation, and 1.3 times the odds of having teachers who describe Unproductive supports for struggling students.

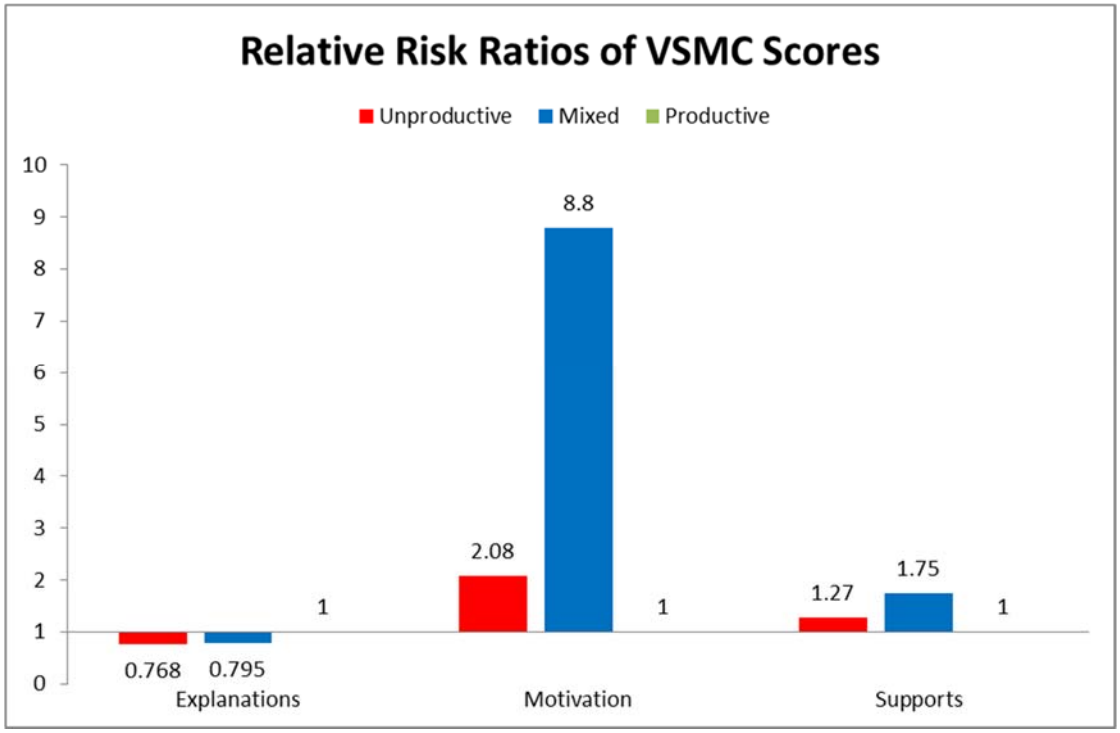

Figure 16:
Relative Risk Ratios of VSMC Scores Predicted from Tracking

An important assumption of the multinomial logit model used for the above estimation is "Independence of Irrelevant Alternatives" (IIA). This means that, if one category were removed or added, the relative probabilities between the other categories would not change. In other words, if the data were re-scored without "Mixed" as a valid

option, then the probability of a teacher being scored as Productive, as compared to their probability of being scored as Unproductive, would not change. One way this data could violate the IIA assumption is if coders were more likely to give teachers a "Mixed" score if they were leaning Unproductive than if they were leaning Productive. Eliminating the Mixed category and forcing coders to choose between Productive and Unproductive would then increase the probability of being scored Unproductive, and likely change the relative probability of the two remaining categories.

To test this possibility in my data, I ran the Hausman IIA test after each of my models. This test examines whether the relative log odds change when excluding each of the categories. In each case I rejected the null of independence, indicating that there is likely to be significant dependence between the categories in this model. There are three options to address this issue. First, I could use an ordinal logit, which assumes that the Unproductive, Mixed and Productive categories follow one after the other in rank order. However, I cannot say with confidence that the categories in these rubrics are ranked because of the nature of the Mixed category. For some teachers, "Mixed" indicates some Productive and some Unproductive statements, while for others it indicates a wavering between views in the same statement. It is not clear if this is a category on a trajectory between Unproductive and Productive, or if it represents a different kind of view altogether. Second, I could use alternative-specific multinomial probit regression. This requires that the independent variable of interest be "alternative specific," or intrinsically tied to the outcome. However, VSMC scores are not intrinsically tied to the outcome, so

my model is not alternative-specific. Therefore, my only remaining option is to examine a binary logistic regression model, comparing only two categories at once[10].

Table 18 shows the outcome of a model using these logistic regressions. First, I compared Productive to Mixed scores, excluding Unproductive scores. These findings are consistent with those above, indicating higher odds of Productive Explanations for why students struggle in tracked settings, but lower odds of Productive Motivations or Supports scores in tracked as compared to untracked settings.

Next, I compared Productive to Unproductive scores, excluding Mixed scores. These findings were consistent with the multinomial logit models on Motivation and Supports, as students in tracked settings were predicted to have higher odds of teachers with Unproductive views on Motivation and Supports. However, the coefficient on Explanations was not statistically significant, indicating that there is not a significant difference between tracked and untracked students in the odds of having a teacher with Productive views as compared to having a teacher with Unproductive views.

[10] I also examined logistic regressions combining the "Mixed" category with either Productive or Unproductive. The results are reported in the Sensitivity Tests section.

Table 18:
Binary Logistic Regression Predicting VSMC Scores from Tracking

	Explanations	Motivation	Supports
Productive vs. Mixed Tracked	0.33**	-2.11***	-0.66***
	(0.05)	(0.11)	(0.06)
Constant	1.42***	-6.21***	4.17***
	(0.30)	(0.47)	(0.46)
Observations	13887	6805	12075
Unproductive vs. Productive Tracked	-0.08	0.67***	0.27***
	(0.09)	(0.14)	(0.08)
Constant	0.25	4.50***	5.68***
	(0.45)	(0.59)	(0.37)
Observations	6949	4081	12888

Standard errors in parentheses
* p < 0.05, ** p < 0.01, *** p < 0.001

Although the initial multinomial logistic regression may suffer from a violation of the basic assumptions of IIA, similar results were found when narrowing the comparison to only two categories. Across the specifications, students in tracked settings are predicted to have lower odds of having a teacher who describes Productive supports for struggling students or has Productive views of student motivation. Students in tracked settings may have slightly higher odds of having teachers with Productive Explanations for why students struggle, but this difference was small, and may only be a distinction between the Productive and Mixed categories rather than the Productive and Unproductive categories.

Research Question 4: Are more productive views of students' mathematical capabilities associated with higher student achievement? There were mixed results on the relationship between teachers' views and student achievement. As shown in Table 19, when teacher controls were not included, students with teachers who had Productive explanations for why students struggle in mathematics were predicted to score about 0.06 standard deviations higher than students whose teachers had Unproductive Explanations. Similarly, students whose teachers had Productive views of student motivation were predicted to score about 0.05 standard deviations higher than those whose teachers had Mixed Motivation scores. On the other hand, Supports scores were not significantly associated with student achievement, controlling for student and school demographics.

When teacher controls (race, number of advanced math courses and years of experience) were added, only teachers' descriptions of supports were significantly associated with student achievement. Students whose teachers described Productive Supports were predicted to score about 0.08 standard deviations higher than those with similar teachers who described Unproductive Supports. The difference between these two models indicates that some of the association between VSMC and student achievement is actually due to differences in the characteristics of the teachers who hold Productive views.

Table 19:
Linear Regressions Predicting Student Achievement from VSMC Scores

	Without Teacher Controls			With Teacher Controls		
	Explanation	Motivation	Support	Exp	Mot	Sup
Productive	0.06*	0.04	0.009	0.03	0.04	0.08**
	(0.03)	(0.03)	(0.02)	(0.04)	(0.05)	(0.03)
Mixed	0.05	-0.01	0.01	0.02	0.01	0.04
	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)
Constant	0.21	0.17	0.28*	0.27	0.02	0.35*
	(0.13)	(0.15)	(0.12)	(0.17)	(0.18)	(0.15)
Productive vs. Mixed	0.01	0.05*	-0.001	0.01	0.03	0.04
Obs	15676	8461	15926	9587	5001	9663

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Adding teacher controls significantly reduced the sample because many teachers did not respond to the survey or did not complete all the demographic items. Therefore, I also imputed missing values of teacher characteristics based on the non-missing values of other variables[11] in the dataset using multivariate normal regression in Stata, which uses an iterative Markov Chain Monte Carlo (MCMC) method to fill in five plausible values based on the values in variables with no missing data (Statacorp, 2009, p. 145). Running the models on imputed data and including teacher controls provided in similar results as those shown in the far right columns of Table 19: a statistically significant relationship

[11] Variables with no missing data that were used to impute the teacher variables were: school size, percent white at the school level, year, grade level, district and average student achievement.

between supports and student achievement, but no significant relationship for Explanations or Motivation.

Research Question 5: To what extent do teachers' views of students' mathematical capabilities moderate the relationship between track level and student achievement? As shown in Table 20, there was a significant moderation effect of Supports on the relationship between tracking and student achievement, but no significant moderation effect of Explanations or Motivation.

Table 20:
Linear Regression Predicting Student Achievement from the Interaction between VSMC and Tracking

	Explanations	Motivation	Supports
Tracked	0.04	0.04	0.15**
	(0.13)	(0.07)	(0.05)
VSMC – Mixed	0.07	-0.06	0.16**
	(0.12)	(0.08)	(0.06)
VSMC – Productive	0.06	-0.02	0.24***
	(0.12)	(0.08)	(0.04)
Tracked * VSMC Mixed	-0.06	0.06	-0.15*
	(0.13)	(0.09)	(0.07)
Tracked * VSMC Productive	-0.03	0.09	-0.22***
	(0.13)	(0.09)	(0.05)
Constant	0.21	0.05	0.15
	(0.21)	(0.17)	(0.16)

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

This is also shown graphically in *Figure 17* and *Figure 18*. In Explanations (*Figure 17*), the gap in student achievement between tracked and untracked settings was approximately the same size, whether the teacher had Productive, Unproductive or Mixed views. One can see this by comparing the two green bars for Productive teachers, the two blue bars for Mixed teachers and the two red bars for Unproductive teachers. Likewise, the gap between having a teacher with Productive and Unproductive Explanations scores (the difference between the green and the red bar) was about the same size in tracked and untracked settings. Although the achievement looks to be the lowest among untracked students whose teachers had Unproductive views, the difference between this bar and the rest was not significantly different than zero.

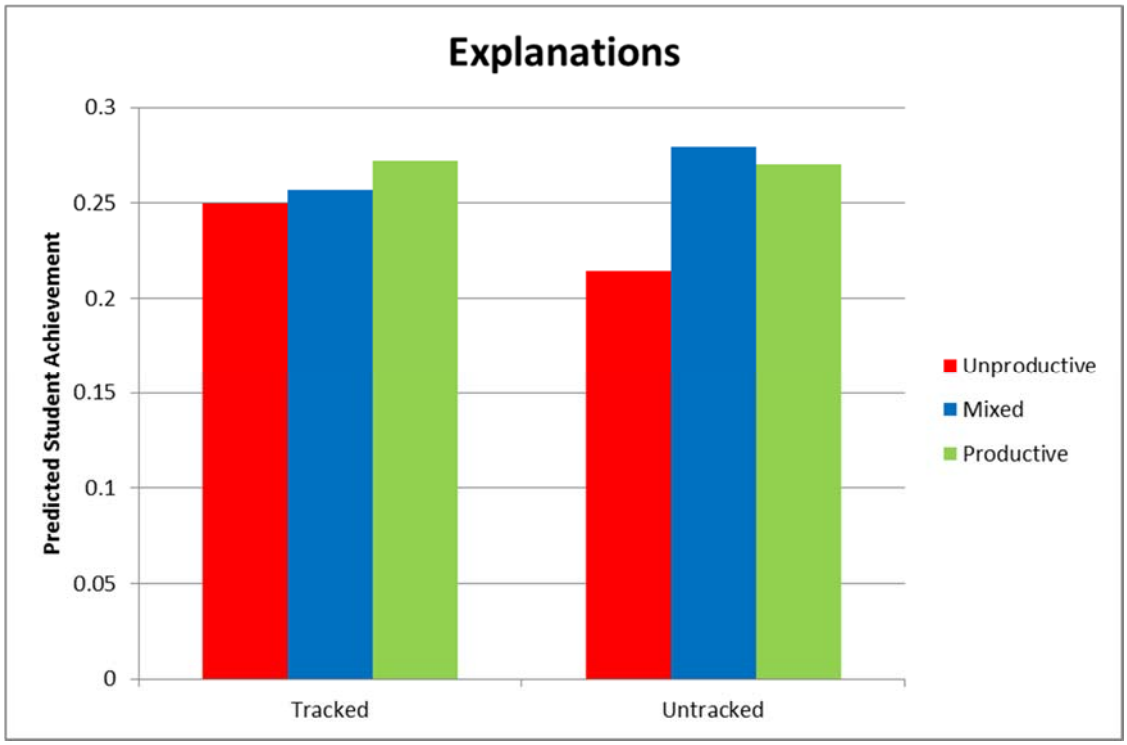

Figure 17:
Effect of the Interaction between Teachers' Explanations for Student Struggle and Tracking on Student Achievement

The relationship between tracking and student achievement was not significantly moderated by teachers' views of student motivation either. In other words, the difference between having a teacher with Productive views of motivation and having a teacher with Unproductive views of motivation was about the same size in tracked and untracked schools.

In contrast, the supports that teachers described were a significant moderator on the relationship between tracking and student achievement (see *Figure 18*). Students whose teachers described Productive Supports for struggling students were predicted to score about 0.06 standard deviations higher in untracked settings than in tracked settings. On the other hand, students whose teachers described Unproductive Supports for struggling students were predicted to score about 0.15 standard deviations *lower* in untracked than in tracked settings. There was no statistically significant difference in tracked settings between having a teacher with Productive views and having a teacher with Unproductive views, but in untracked settings, this gap was nearly one-fourth of a standard deviation. This indicates that having a teacher who describes supports for struggling students that engage them in rigorous mathematics is more important in untracked than in tracked settings, and that students in untracked settings actually outperform those in tracked settings if their teachers describe such supports.

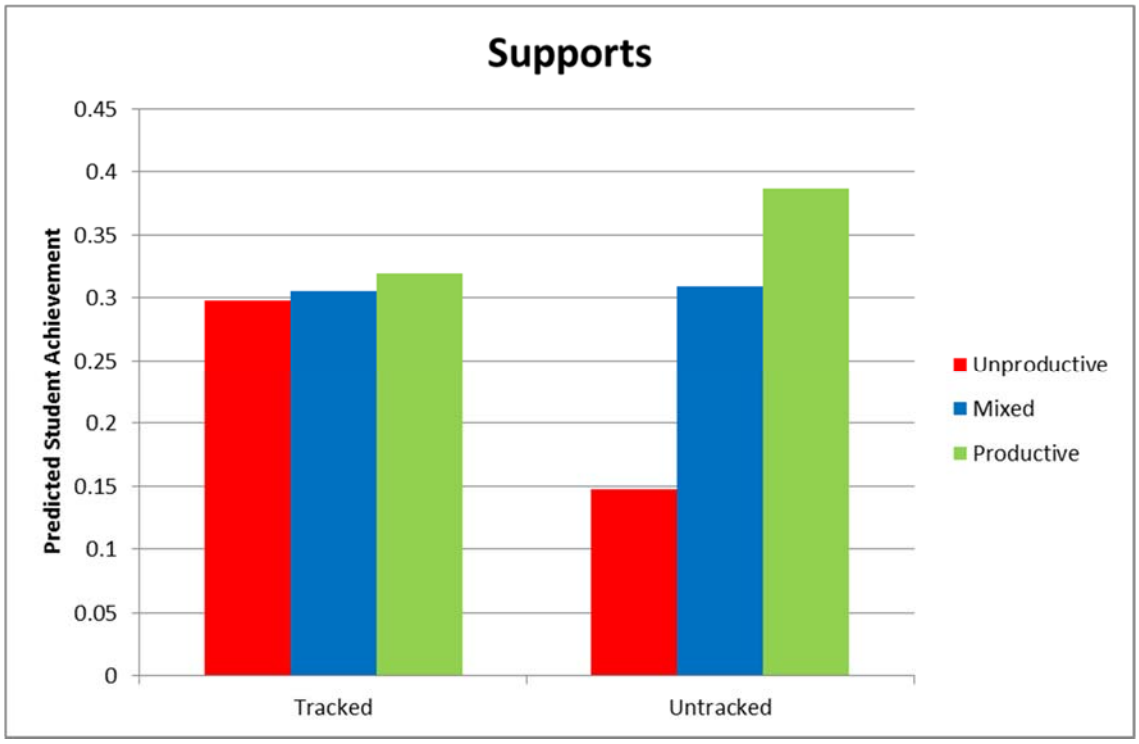

Figure 18:
Impact of the Interaction between Teachers' Views of Supports for Struggling Students and Tracking on Student Achievement

As with the models used in the previous research question, including teacher control variables greatly reduces the sample. Therefore, I used the same multiply-imputed dataset to estimate these models again on a larger sample. Using this sample, the interaction between tracking and teachers' Explanations for why students struggle was statistically significant. This aligns well with what was shown graphically in *Figure 17* but was not statistically significant in the smaller sample. In the imputed dataset, the difference in achievement in tracked settings between students whose teachers had Productive views and those who had Unproductive views was still statistically insignificant, but the difference in untracked settings was about 0.1 standard deviations. Untracked students whose teachers had Productive Explanations were not predicted to

outscore tracked students, but untracked students whose teachers had Unproductive views were predicted to score significantly *lower* than tracked students. In other words, having a teacher who explains student struggle as something that can be influenced by instruction matters more in untracked than in tracked settings.

Therefore, both the way teachers explain students' struggle and the way they support students in mathematics have a significant impact on the relationship between tracking and student achievement. Students in untracked settings whose teachers describe Productive supports significantly out-score their counterparts in tracked settings, indicating that untracked settings can be associated with higher achievement if they are paired with teachers who believe in supporting students to engage in rigorous mathematics. While Productive Explanations scores are not associated with higher achievement among untracked students, *Unproductive* Explanations scores are associated with significantly lower achievement. This means that untracked settings can be associated with *lower* achievement when they are paired with teachers who believe that students struggle for reasons that they (the teacher) cannot influence or correct.

Sensitivity Tests

Combining Categories. The results of the Hausman test for the Independence of Irrelevant Alternatives in the first three research questions led me to substitute logistic regression for the multinomial logit that makes use of all three categories. To do this, I had to run separate models comparing Productive to Mixed and Productive to Unproductive. However, another possible approach is to combine the Productive and Mixed category and compare these to the Unproductive category (Unproductive versus not Unproductive), or, alternatively, to combine the Unproductive and Mixed categories

and compare them to the Productive category (Productive versus not Productive). As shown in Table 21, the specification combining Mixed and Unproductive into one category (Model 1) was consistent with both the multinomial logit and the logistic regression findings above: students in tracked settings had a significantly higher likelihood of having teachers with Productive Explanations of why students struggle, but a lower likelihood of having teachers with Productive views of student motivation or teachers who described Productive supports for struggling students.

Combining Productive and Mixed into one category (Model 2) yielded different results. In this specification, students in tracked settings had a significantly lower likelihood of having teachers with Unproductive views of student motivation, but no significant difference in Explanations and Motivation. This seems to indicate that the "Mixed" category is serving a functional purpose: there is some difference in the probability not just of having a teacher with Productive versus Unproductive views, but also in the probability of having a teacher with Mixed views. Although I cannot establish that Mixed is between Unproductive and Productive on an ordinal scale, this set of models indicates that it does function as a separate category, not a subset of either Productive or Unproductive. This is borne out by Model 3, which combines Productive and Unproductive into one category and compares them to Mixed views. The reason for this comparison is not because I believe the Productive and Unproductive categories should be combined, but just to test whether the conclusions found in the multinomial logit are supported when looking at each category separately. In this specification, tracked students are significantly less likely to have teachers with Mixed Explanations of student performance, but more likely to have teachers with Mixed views of Motivation

and Supports. Therefore, despite the violation of the IIA assumption, the findings using multinomial logistic regressions are supported by logistic regression, regardless of the way the categories are combined.

Table 21:
Binary Logistic Regression Predicting VSMC from Tracking using Different Combinations of Categories

	Model 1: Productive vs Not Productive	Model 2: Unproductive vs Not Unproductive	Model 3: Mixed vs Not Mixed
Explanations	0.29***	-0.13	-0.12**
	(0.05)	(0.07)	(0.05)
Motivation	-1.56***	-0.63***	1.84***
	(0.10)	(0.10)	(0.09)
Supports	-0.46***	-0.03	0.44***
	(0.05)	(0.07)	(0.06)
Observations			
Explanations	15937	15937	15937
Motivation	8618	8618	8618
Supports	16183	16183	16183

Stable Sample across Rubrics. Each teacher in this sample was rated on at least one VSMC rubric. However, as indicated by the number of observations in each table, teachers could be rated for Explanations, but not Motivation, or for Supports but not Explanations. Therefore, the conclusions about each aspect of teachers' views are based on a slightly different sample. To test whether this difference in sample affected my results, I re-ran the models for each research question, restricting the sample to only those

teachers who had scores on all three rubrics. This reduced the sample to 55 unique teachers (between 11 and 20 in each year) and their 5,231 students across the four years.

As shown in Table 22, students in tracked settings were still predicted to have lower probabilities of having teachers with Mixed Explanations scores and higher probabilities of having teachers with Mixed Supports scores. However, the coefficients on Explanations and Supports were both statistically insignificant for the Unproductive category in this reduced sample. Additionally, the coefficients on Motivation switched sign, so that students in tracked settings were predicted to have a *lower* probability of having teachers with Unproductive or Mixed Motivation scores. This indicates that the relationship between tracking and teacher beliefs may be different among teachers who were scored on all three rubrics. However, the reasons why some teachers may be scored on all three rubrics rather than only one or two are not clear. There were no systematic relationships between any of the variables used in this analysis and the odds of being scored on all three rubrics, except for grade level. Eighth grade teachers were significantly more likely to have scores on all three rubrics. It is possible that this indicates that eighth grade teachers are more explicit in describing their views, so that it is easier to code their interviews. However, the main analysis does control for grade level. Absent a more satisfying explanation for this finding, the reduction in sample size is the most plausible reason that the tracking was found to be insignificant in predicting Explanations and Supports rubrics.

Table 22:
Research Questions 1 – 3 with a Stable Sample across Models

	Explanations	**Motivation**	**Supports**
Unproductive			
Tracked	-4.23	-4.77***	-16.32
	(1905.80)	(0.28)	(532.78)
Constant	-39.55	-21.29	-8.60***
	(2306.74)	(599.71)	(0.85)
Mixed			
Tracked	-5.65***	-1.40***	1.83***
	(0.36)	(0.22)	(0.22)
Constant	-4.94***	0.23	-9.72***
	(0.92)	(0.74)	(0.83)
Observations	3802	3802	3802

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In the fourth research question, whether there is a relationship between VSMC scores and student achievement, narrowing the sample to teachers who had scores on all three rubrics resulted in statistical insignificance in all three models. This is also probably due to the reduction in sample size. In the fifth research question, whether there is a significant moderating effect of VSMC on the relationship between tracking and student achievement, the coefficients in the Explanations and Motivation models were both statistically insignificant, but the coefficient in the Supports model remained significant and in the expected direction. In other words, reducing the sample did not affect the conclusion that having a teacher who discusses Productive Supports matters more in untracked than in tracked settings.

This sensitivity test indicates that some of the differences between rubrics in the relationship between VSMC scores and tracking may have been due to a different sample of teachers scored on these items. The reasons why some teachers do not have scores on

all three rubrics are because of a lack of sufficient information in the interviews or a lack of reliability in the coding. Coding reliability was checked consistently and remained at over 80% throughout the coding process, so this is unlikely to have caused the difference. Insufficient information in the interviews arose because key questions were not asked, the interviewer did not press for clarification, or the teacher did not adequately explicate his or her views. The first two are likely to be randomly distributed among teachers, but the last could indicate one reason for the different findings in this sensitivity test. If teachers who are more able to verbalize their views are systematically different from those that could not, then narrowing the sample to only these teachers could find different results. Therefore, this sensitivity test suggests two solutions. First, using teacher-level fixed effects to examine each research questions within rather than between teachers. This will be addressed below. Second, there is a need for further research using data where all teachers were adequately pressed to elaborate their views, to allay any potential response bias.

Multi-level Modeling. The models used above to answer the fourth and fifth research questions employed clustered standard errors to account for the fact that student achievement is likely to be more similar within than between classrooms. However, teachers may also be clustered within schools, which can be addressed use multi-level modeling. In these two research questions there are potentially five levels: students nested within years nested within teachers nested within schools nested within districts. However, I already used district fixed effects, so I did not include this as a level. When I examined unconditional models, there was significant variation at the school level (intraclass correlation, or ICC, of 0.12), and moderate variation at both the teacher level

(ICC of 0.04) and the year level (ICC of 0.06). Therefore, I re-analyzed research questions four and five using multilevel modeling to account for the nested structure of the data. As expected, this did not impact the size of the coefficients in either case, and only moderately affected the standard errors. Therefore, the conclusions drawn above using clustering only at the classroom level are supported using multilevel modeling, but may slightly underestimate the standard errors.

Peer Effects. One potential rival explanation for the relationships found here between teachers' views of student ability, tracking and student achievement is peer effects. If there is a correlation between the average prior achievement of students, their enrollment in an untracked school and teachers' views of their ability, then not controlling for this variable above could lead to omitted variable bias. What appears to be a relationship between teachers' views and achievement could in fact be the effect of students' peers. To test this, I added classroom average prior achievement to each of the models in all five research questions. The addition of this variable did not change the coefficients in the models testing research questions 2 or 3, indicating that the relationship between tracking and teachers' views of motivation and supports cannot be attributed to differences in the prior achievement of the students they are exposed to. However, controlling for students' prior achievement resulted in insignificant coefficients on the relationship between teachers' explanations for student struggle and tracking. This suggests that students in tracked settings were only more likely to have teachers with productive explanations for students' struggle when their classrooms also had higher average achievement.

Introducing peer effects to the models for research questions 4 and 5 did not affect the size or the significance of the coefficients. This shows that the relationship between productive views of supports for struggling students and student achievement could not be explained by peer effects. Likewise, having a teacher with Productive views matters more in untracked settings, even controlling for the average prior achievement of the class. Hence, most of the relationships found above could not be attributed to the effect of the students' peers.

School and Teacher Fixed Effects. As mentioned in the Methods section, there are many teacher and school factors that may be correlated with teachers' views of students' mathematical capabilities, tracking and student achievement. This analysis attempted to control for as many of these factors as possible, but to the extent that there is non-random sorting of teachers between tracked and untracked settings, there still may be unobserved variables that could not be controlled. Therefore, I examined the final research question using teacher- and school-level fixed effects.

Using school fixed effects, the interactions between tracking and both the Explanations and Support rubrics were statistically significant, just as found above in the models using the imputed dataset. This indicates that, within schools, having a teacher with Productive Explanations or Supports scores mattered more in untracked than in tracked grade levels. In untracked grade levels, the gap in student achievement between students whose teachers had Productive Explanations and those whose teachers had Unproductive Explanations was about 0.2 standard deviations, while the same gap in tracked grade levels was not statistically significantly different than zero. Likewise, in untracked grade levels, students whose teachers had Productive descriptions of supports

for struggling students were predicted to score about 0.1 standard deviations higher than students whose teachers had Unproductive Supports scores. There was not a statistically significant difference in tracked settings.

Using teacher-level fixed effects, the interaction between Explanations and tracking was not statistically significant, but the interaction between Supports and tracking remains statistically significant ($p<0.05$). Teacher fixed effects examine the relationship "within teachers." In this case, the model looks at teachers who either moved from tracked to untracked settings (or vice versa) or changed in their Views of Students' Mathematical Capabilities across time, or both. Therefore, a teacher who describes Productive Supports for struggling students is predicted to have higher student achievement when they teach untracked than when they teach tracked students. Likewise, if a teacher remains in an untracked setting, but moves from describing Unproductive to describing Productive supports, their students are predicted to have higher achievement, even after accounting for their prior achievement. As shown in *Figure 19*, the size of this relationship is about the same as without teacher fixed effects.

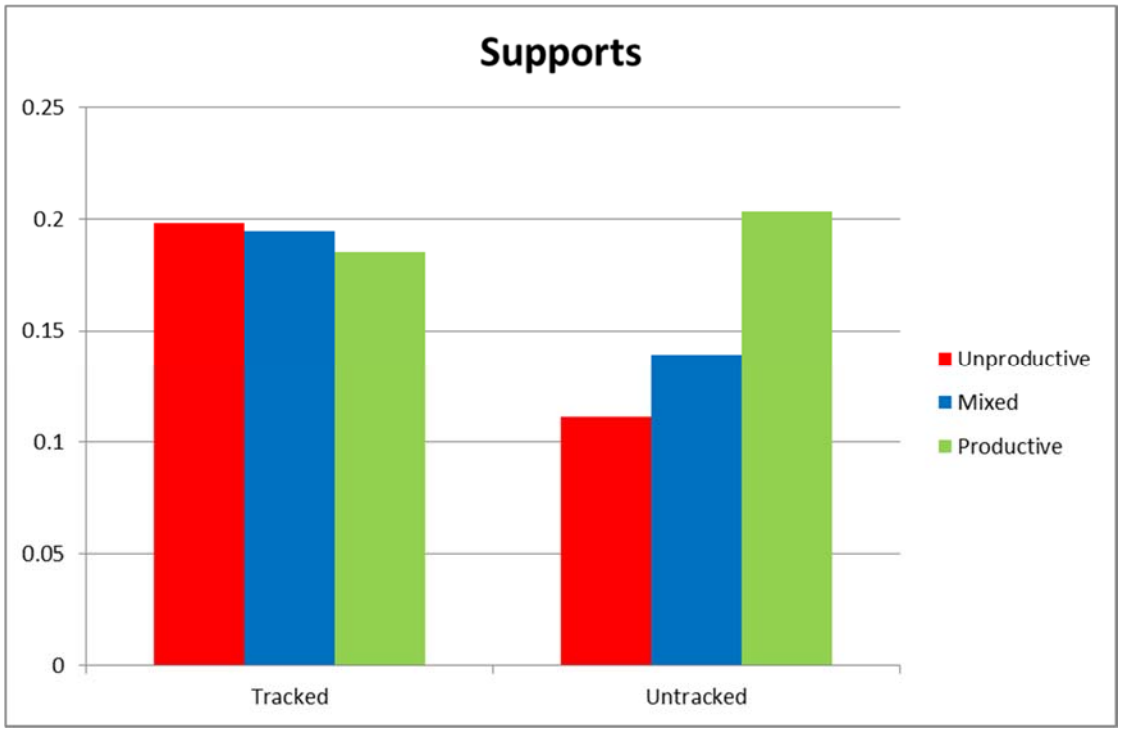

Figure 19:
Teacher Fixed Effects Estimation of the Impact of the Interaction between Teachers' Views of Supports for Struggling Students and Tracking on Student Achievement

This sensitivity test shows that the moderating effect of Productive views of Supports for struggling students is robust to the non-random sorting of teachers. In other words, it is not unobservable differences between teachers or schools that can explain the effect, strengthening the claim that students in untracked settings can out-perform those in tracked settings if their teachers hold Productive views of Supports for struggling students.

Limitations

There are two main limitations of this analysis. The first is the problem of non-random sorting of students, and the second has to do with the measure of teachers' views

of students' mathematical capabilities. The non-random sorting of students could affect my analysis if there are unobservable differences between tracked and untracked students that are correlated with both achievement and their teachers' VSMC scores. Because the reasons a school would be untracked often have to do with the beliefs of the school administrators and the community, it is conceivable that these views as well as the students' own beliefs about their abilities could be correlated with both their teachers' views and their own achievement. School-level fixed effects would account for administrator and school beliefs, but not for students' own beliefs. I attempt to control for any pre-existing differences between students using their prior achievement test scores and demographic variables, but this may not capture all of the effect.

Unfortunately, the data used in this analysis do not allow for student-level fixed effects. Although the districts provided prior achievement for each student, they did not provide information on the school or classroom the student was in the previous year, so I cannot examine whether moving from a tracked to an untracked setting or from a teacher with Productive to a teacher with Unproductive views has the same impact as the results found here. Therefore, the results found here may incorporate both the impact of teachers' views about students and some unobserved differences between students, such as their own beliefs about themselves.

The second limitation in this analysis is the measure used of teachers' views about student ability. This measure advances our previous understanding of the relationship between teachers' views of ability, tracking and student achievement by beginning with a specific set of beliefs that need to be measured and quantifying the *size* of this relationship. However, because the interview was not designed for this measure, many

teachers did not have scores on all three rubrics. In fact, only eleven to twenty teachers in each year had scores on all three, 36 to 45 had scores on two (usually Supports and Explanations), and 33 to 56 had scores on only one rubric (usually Supports or Explanations). This is in part because the interview questions discussed in the Measures section changed from year to year, and some of the more valuable questions were not always asked by each interviewer. As a result, the models used to answer these research questions use a different sample of teachers for each rubric. I attempted to address this in the Sensitivity Tests section by running the models on only the sample of teachers that had scores on all three rubrics, but this reduced the number of observations dramatically, and only the interaction between Supports and tracking remained significant. Additional analyses using a quantitative measure of teachers' beliefs should apply this measure to interviews focused on this topic and with interviewers trained to press teachers to clarify their views.

Relatedly, the "Mixed" category in each VSMC rubric did not have a clear interpretation. While Productive views are aligned with a developmental conception of ability and Unproductive views support a more static view, "Mixed" views can include teachers who alternate between views when talking about different groups of students or those who say both Productive and Unproductive things, even when talking about the same students. Teachers with Mixed views may be transitioning between Unproductive and Productive views, but it may also be that the interviewer did not ask enough clarifying questions to help a teacher explain their beliefs about students, or that the teacher holds a different kind of view altogether. The coding approach did not distinguish between teachers who said mostly Unproductive things, with only one or two

Productive responses, and teachers whose responses were the opposite. Future analysis could use the same rubrics but attempt to break down the Mixed category in this way.

Conclusion

Recent researchers have argued that tracking is inequitable because it relies on a definition of ability as innate, uni-dimensional and unchanging. These researchers advocate instead for a developmental definition of ability: one in which students' abilities may change over time and may be influenced by instruction. Tracking, under this view, cannot be efficient or equitable because it assumes that ability is stable at least from one test to the next, and it separates students on the basis of this assumption. Even while arguing for detracking, however, many researchers recognize that teachers' own beliefs about the nature of ability may constitute a barrier to successful implementation of untracked classrooms. If teachers believe that ability is innate and cannot be impacted by their instruction, then they are likely to continue to sort students within their classroom and to categorize students as "low" and "high" ability. Therefore, a developmental conception of ability may be necessary to successful detracking.

Although prior qualitative research has found an association between teachers' views about the nature of student ability and both tracking and their students' achievement outcomes, these studies have not started with a specific measure of the aspects of teachers' views expected to matter or linked the three quantitatively. This paper uses a quantitative measure of teachers' views in three categories: explanations of students' performance, views of students' motivation, and the nature of instructional supports for struggling students. "Productive" views on each of these rubrics align with a

developmental conception of ability, which many researchers argue is necessary for the success of detracking initiatives.

This analysis found that students in tracked settings had about twice the odds of having teachers with Unproductive views of student motivation, and 1.2 times the odds of having teachers who describe Unproductive supports for struggling students. On the other hand, students in tracked settings had *lower* odds of having teachers with Unproductive explanations for why students struggle or succeed in mathematics. The Supports rubric was also significantly associated with student achievement and had a significant moderation effect on the relationship between tracking and student achievement. This means that teachers' views of supports for struggling students mattered more in terms of student achievement in untracked than in tracked settings. Students in untracked settings were predicted to out-score their counterparts in tracked settings if their teachers described Productive Supports for struggling students. The significant moderation effect of Supports remained significant through all specifications of the model. This finding supports the contention that the success of detracking efforts may depend on teachers buying into a developmental definition of ability.

This study provides a point of entry for improving the implementation of detracking initiatives: supporting teachers in understanding and incorporating a developmental conception of ability. While changing teachers' beliefs may be difficult, it is clear that how they conceive of their students interacts with ability grouping policies to have a significant effect on student achievement. Therefore, policies must attend to these beliefs and help teachers support all students to participate in rigorous mathematics.

CHAPTER IV

"DOUBLE DOSE" POLICIES AS A SUPPORT FOR LOW-ACHIEVING STUDENT IN MATHEMATICS: CHARACTERISTICS AND RELATIONSHIP WITH STUDENT ACHIEVEMENT

In addition to teachers' beliefs and practices, school-level policy decisions may affect students' achievement in both tracked and untracked settings. Students enter middle school in the United States at disparate achievement levels and it is often up to the schools to address this issue. Ability grouping at the classroom level, or tracking, is a school response with a long history and much debate over its success. While tracking has diminished in recent decades, the sizeable number of struggling students in middle schools often makes classroom-level interventions more attractive than individual-level interventions. One newer classroom-level intervention that has little research support is "double dose" classes in core subjects such as mathematics. Double dose mathematics classes provide an additional period of mathematics for struggling students while also providing them with a "regular" mathematics period. The provision of significant instructional time is associated with improved student achievement (e.g., Baker, Fabrega, Galindo & Mishook, 2004; Bloom, 1974; Brown and Saks, 1986; Silva, 2007), but the particular programmatic choices of double dose have been understudied.

While double dose policies always provide a substantial increase in instructional time, schools must make many other implementation choices, most of which have been overlooked by prior research. First, double dose instruction may take place in the context of tracking, so that low-achieving students are grouped into separate classrooms and also

provided with an additional period of mathematics instruction. On the other hand, it may take place in untracked schools, where low-achieving and high-achieving students share their "regular" mathematics class, but low-achieving students receive an additional period of math instruction (Nomi & Allensworth, 2011; Wheelock, 1992). Double dose classes may also target the lowest-achieving students, or they may target "bubble kids": those whose scores fall within a range that is close to, but just below, passing the state tests (Booher-Jennings, 2005; Perkins-Gough, 2006). Third, what is taught in these classes can vary a great deal. Some classes have no set curriculum, while others explicitly mirror what is taking place in "regular" instruction (Nomi & Allensworth, 2011). Finally, the teachers who lead double dose classes may be the same as those in regular classes, or there may be a designated "double dose" teacher (Nomi and Allensworth, 2011).

This study advances the research on double dose instruction by not only examining its association with student achievement, but also assessing how the impact varies for different student populations and under different conditions. Through interviews with teachers and principals in middle school mathematics in four large, urban districts, I examine how the face of double dose varies from school to school. Then, using student achievement data from these same districts, I test the relationship between the presence of double dose and student achievement for both students in double dose classes and the non-double-dose students who attend the same schools. Finally, I ask whether these "double dose effects" vary by the programmatic components of double dose policies, asking whether it is simply the additional time that affects student achievement, or if it matters how that time is organized.

Literature Review

As districts have moved away from classroom-level ability grouping, several researchers have drawn attention to the continued disparity in incoming student achievement and insisted that detracking initiatives must be accompanied by other supports for low-achieving students. Nonetheless, few studies have investigated the nature of the supports put in place or how effective they may be at improving students' achievement, particularly in mathematics. In fact, What Works Clearinghouse (WWC) reports very few studies examining supplemental supports for low-achieving students in mathematics, regardless of the tracking context. Thirty-three studies on mathematics education met WWC's evidence standards as of April 2012. Only five studies showed at least potentially positive effects for use with middle- or high-school students. All five were either inquiry-based curricula or interactive software, and only one (Cognitive Tutor) was designed for use with struggling students outside the context of the regular classroom.

Additional Instructional Time. One way schools can support students who enter middle school at lower achievement levels is to provide them with additional instructional time. Bloom (1974) made the seemingly straightforward argument that learning takes time, and time for everyone is limited, but time for learning in school is particularly limited. The gap between the highest and lowest state in high school achievement at the time of Bloom's study was about one standard deviation, which was the same as the gap between the highest and lowest developed nations. Students in the lowest-achieving states and nations completed about an 8th grade education compared to a 12th grade education in the highest states and nations. The author argued that this was

evidence that students learn at different rates because of different inputs in terms of time and resources. Bloom also argued that under the right conditions, the vast majority of students could learn at the same level as the highest achieving. The "right" conditions require extra time for students who fall behind, usually outside of the classroom. Bloom argued that if this additional time is provided during the school year, rather than the summer, it provides "psychic and motivational rewards" (p. 683) in addition to the learning rewards.

Since Bloom's (1974) seminal work, a number of researchers have examined the relationship between instructional time and student learning. Brown and Saks (1986) looked at second and fifth grade classrooms and found a significant positive association between allocated time and achievement in both reading and math. These effects varied by classroom in mathematics, but not in reading. Likewise, Marcotte (2007) found that when snowfall reduced the number of school days experienced before the state assessment, students had significantly lower test scores than in years that were unaffected by winter weather. The author found that this effect was stronger in mathematics, possibly due to "relatively inflexible curricula" (p. 629) common to math classes.

Schools and districts commonly introduce greater instructional time through "block scheduling." These policies double the amount of instructional time for *all* students in a school, district or grade, creating mathematics courses that run for a full 80

to 90 minutes every day[12]. These policies differ from "double dose," which targets only struggling students, but their impact may suggest the effect of increased instructional time in practice. Peele (1998) examined the outcomes of an experimental block scheduling policy in one school in Norfolk, Virginia, during the 1995-96 school year. The school randomly assigned students to either the control condition (a one period mathematics course) or the experimental condition (a double block scheduled course). In this study, the block scheduled course was two separate periods, in which the second class was taught by a different teacher and designed to reinforce what was learned in the first class, using computer software, homework help and other curricular materials at the teacher's discretion. The study was very small, with only two teachers participating, each of whom taught one period of the block scheduled course and one control course. The authors found no statistically significant differences between treatment and control (likely due to the very small sample size). Descriptively however, grades on the final exam were 5 points higher among block scheduled students, passing rates were 4.2 percentage points higher, and there were more As and Bs and fewer Cs and Ds among block scheduled than control students.

The Talent Development High School model also employs increased instructional time. Balfantz, Letgers and Jordan (2004) evaluated the effectiveness of this program in three nonselective high schools in Baltimore. These schools were matched on demographics and prior achievement to three comparison schools. The Talent

[12] This is usually called "double block" scheduling. In "single block" scheduling, classes meet for 80 to 90 minutes *every other day*, which does not increase the total amount of instructional time, only the amount of time dedicated in one session. "Double block" therefore is parallel to double dose instruction, but for all students.

Development High Schools provided 90 minutes of math and reading instruction to all 9th grade students every day, as well as additional support courses, and professional development for teachers. The curriculum in the block courses emphasized conceptual understanding over test preparation. While implementation of the program as a whole varied, the authors found that students in experimental schools significantly out-performed those in control schools in both reading and math. The effect size for the program in mathematics was 0.18, which was about the same as the effect size of improved daily attendance found in other studies. Finally, despite implementation problems, teachers rated the program highly and said it allowed them to teach more effectively. Similar effects were found in Philadelphia (Balfantz, Letgers and Jordan, 2004; Kemple, Herlihy and Smith, 2005). While this provides suggestive evidence for the impact of programs increasing instructional time, these evaluations do not separate the effect of block scheduling from the other components of the Talent Development High School model, such as professional development for teachers, the ninth grade "success academy" or the additional high school transition courses.

While the above studies found positive effects of increased instructional time overall, some researchers have suggested that increased time may matter more under some conditions than others. Silva (2007) showed that extended time for low-income and minority students may matter more than for affluent students because of differences in out-of-school experiences. Likewise, Brown and Saks' study (above) found that the effects of time allocated to learning were stronger for students with lower prior achievement. Internationally, Baker, Fabrega, Galindo and Mishook (2004) examined the relationship between instructional hours and achievement outcomes in three cross-

national datasets and found that the only sizable effects were for large differences in instructional time. Similarly, a program to double the length of the school day in Chile had significant impact on student achievement in both mathematics and language arts (Bellei, 2009).

These studies support the contention that instructional time has a significant impact on student achievement, but that this impact may vary by prior achievement and subject, and the increases in time need to be substantial. Therefore, policies providing significant extra instructional time to low-achieving students in mathematics, such as double dose, are of particular interest, and examining the achievement outcomes of such policies is paramount.

Programs and policies that provide additional instructional time. Several studies have examined policies providing additional instructional time to students (e.g., Perkins-Gough, 2006). However, many of these policies do not support only low-achieving students in mathematics, nor do they provide significant additional time—the factors which the studies above indicated may have a larger impact. While double dose classes are aimed specifically at low-achieving students and double the amount of instructional time, some other policies increase instructional time for *all* students (e.g., double block scheduling), provide increases in instructional time that are isolated from the regular school day (e.g., summer school), and/or provide the extra time on a purely optional basis (e.g., tutoring). Research on these types of interventions is much more common than on double dose instruction.

The Council of the Great City Schools (2009) released a report outlining the types of supports provided in 53 districts across the U.S. in October 2007. About 13% of

districts were using the Talent Development model at some of their schools, and a large number were using other forms of additional instructional time: 91% used block scheduling, 26% had extended their school day, and 19% had extended the school year. When asked about their top three reforms, 28% of districts mentioned "double periods of math instruction" and 21% said block scheduling. While "double periods" may be the same as double dose, it could be additional instructional time for all students. On average, 31% of entering 9th grade students in these districts received remedial math instruction or interventions. However, the choice of interventions was made at the school level in 68% of districts. Double dose (49%), after school/summer school (29%) and specialized math courses (9%) were the most common interventions. Very few of these interventions were being evaluated either formally or informally. In fact, 31% of districts conducted no evaluation (internal or external, formal or informal) of their math interventions.

Chait, Muller, Goldware and Housman (2007) also described various interventions for low-achieving students, including a section on "extended learning time programs." These programs came in a variety of forms, from after school and summer programs, to additional time during the school day. Examples include "shadow classes," designed to take place directly after core courses and reinforce the skills learned during those classes, and double block scheduling, which simply doubles the amount of time spent in core classes. The authors cite two examples of extended learning time programs currently underway: Massachusetts has lengthened the school day for middle school students state-wide, and Louisiana has provided a double blocked, accelerated curriculum to 30 high schools. Although both of these programs provide additional instructional

time, neither focuses primarily on low-achieving students, and the outcomes of the studies have not yet been published.

Mac Iver (1991) reviewed remedial programs for struggling students as reported in the National Education Longitudinal Study (NELS) and Hopkins Enhancement survey. The author found that 96% of schools offered at least one remedial program for struggling students, and the most common remedial activities included extra work, pull-out programs, outside-school tutoring and summer school. Seventeen percent of schools used extra periods in core subjects as an intervention. This intervention was significantly more common in secular private schools than in other school types. The authors also investigated the effect of remedial programs on math achievement. In their analysis, the authors looked only at students who replied that they had participated in a "remedial math class at least once a week" (p. 4) and controlled for student characteristics. They found significantly higher achievement among students who attended schools offering an additional subject period instead of an elective (0.15 standard deviations). However, individual students were not linked to their participation in particular programs, only to their enrollment in schools offering these programs. Therefore, the connection between additional time and student achievement could not be adequately established in this study.

Additional support classes for struggling students: True "Double Dose."

Although there are indications that many districts are employing additional support classes for struggling students, there have been few evaluations of their impact on student achievement (Council of the Great City Schools, 2009). In a study in one Texas district, Cavanagh (2006) found that achievement in the district increased after they introduced

"double dose" classes (additional instructional time) for low-achieving students, although this is a correlational result: the author did not establish that the increase in achievement was not caused by other factors, nor did he investigate whether the increase was stronger among those who received double dose instruction than among those who did not.

Burris, Wiley, Welner and Murphy (2008) examined a suburban district in Long Island, New York that gradually detracked and provided a challenging mathematics curriculum to all middle school students. As a part of this detracking initiative, the district also established extra support classes for struggling students. The authors compared three cohorts of students before detracking to three cohorts of students after the reforms, using logistic regression to predict the attainment of a Regents diploma. Controlling for demographic characteristics and "scholastic aptitude" (a combination of verbal and math PSAT scores), the study found that "detracked cohorts have odds of Regents diploma attainment nearly six times greater than their tracked counterparts" (p. 591). The authors also found that detracking was associated with a *decrease* in the dropout rate, and that the increase in Regents diplomas in this district was larger than the increase in the state overall. Although this study focused on the impact of detracking, the use of additional instructional time for struggling students supports Loveless' (1999) and Gamoran's (2009) contention that detracking can be successful with appropriate supports.

I am aware of only two studies directly examining the impact of significantly increasing instructional time for struggling students. In an unpublished doctoral thesis, Ney (2010) examined the impact of a double-dose policy in one New Jersey high school. Beginning in the 2004-2005 school year, the policy doubled instructional time for students who entered high school without having passed the eighth grade proficiency test.

The paper compared students who received double dose instruction for one, two or three years under this policy (cohorts 2, 3 and 4) to students who received no double dose instruction (cohort 1). He also made these comparisons by race and socio-economic status. In mathematics, Ney found that there was a significant increase in mean scores between cohort 1 (the control cohort) and cohort 2 (the group that had only one year of double dose classes). However, there was a significant *decrease* in mean scores between the control cohort and cohort 4 (the group who had the full three years of double dose classes). A similar pattern of results was found for Hispanic students, White students and non-low-income students. However, Ney did not compare the results for double dose students (positive or negative) to the changes in mean achievement of non-low-achieving students after the policy was in place. The author pointed out that there were some changes in the non-low-achieving students' mean scores as well, but he only rarely stated what those changes were, and provided no test of whether those changes were significantly different from the changes in the "treatment" group. In other words, it is not clear that changes in achievement for students affected by the policy were not mirrored by changes in achievement for students who were not affected by the policy. Additionally, he did not describe the characteristics of this double dose policy, such as the focus of the classes or the teachers used.

The Chicago Consortium on School Research carried out the only large scale evaluation of double dose polices I found (Nomi and Allensworth, 2008, 2011). In 2008 the authors examined the outcomes of a policy in Chicago Public Schools (CPS) to provide a second algebra class for low-achieving ninth grade students (called "double dose Algebra"). This policy was developed in response to high failure rates after the

"College Preparatory Curriculum for All" policy required all students to take Algebra I in ninth grade. Although CPS put forth a preferred model for the implementation of this policy (a regular algebra class plus an additional support class; the same teacher and students in both courses, and the courses offered sequentially), scheduling difficulties meant that many schools did not fully implement this model. Particularly, the majority of schools sorted above- and below-norm students into separate algebra courses (i.e., they tracked students in mathematics), which changed the average ability level of *all* students' classmates. Therefore, the policy, which was intended to impact only below-norm students, may have also impacted above-norm students through differences in classroom composition. The authors used a combination of methods to examine the impact of the policy on both groups of students. First, they examined the Intent to Treat (ITT) effect using a combination of regression discontinuity and interrupted time series design. In this analysis, they found that test scores increased overall post-policy. Second, the authors looked at the differential impact by prior achievement. They found that post-policy, high-achieving students' grades went down and failure rates went up in all percentile groups, though their test scores increased. On the other hand, low-achieving students' grades went up and failure rates did not change. Test scores improved for students at all prior achievement levels, with the largest improvement among students between the 20th and the 50th percentile. Overall, the policy had positive effects on student learning when measured by test scores, but detrimentally impacted the grades of students in higher achievement groups. The authors hypothesized that this could happen because of increased rigor, higher expectations or changes in classroom composition (the "big fish, small pond" effect).

In a second report, released in 2011, the same authors investigated classroom composition further. In this report, the authors clarified that schools chose one of two routes to implementing "double dose" algebra courses: the majority of schools chose to sort students by skill level and provide high-skill students with their own algebra course and low-skill students with a separate, two-period course. This led to more homogeneous math courses in 9th grade in these schools. However, a minority of schools continued to use mixed-ability classrooms, but provided an additional period of mathematics as a separate course for the low-achieving students. The authors examined two cohorts of first time 9th grade students post policy and compared them to cohorts of students prior to the policy. They found that test scores among both above- and below-norm students improved (just as in the 2008 study). Controlling for the prior achievement of students' classmates explained 25% of the improvement in test scores among high achieving students in homogeneous classes, but did not explain any of the improvement for low-achieving students. This indicates that the improvement in test scores among high-achieving students in homogeneous classes may have been because their classes could move more quickly or cover more material without the low-achieving students in them, or because of peer effects. On the other hand, the improvement in test scores of low-achieving students could not be attributed to being grouped with other low-achieving students, and so likely could be attributed to the additional math classes.

Passing rates among above-norm students declined, while passing rates among below-norm students improved under the double dose policy. The authors argued that this may be because students who were just above norms, and so were placed in "regular" classes without double dose, struggled as the difficulty of these classes increased, while

those who were just below norms excelled when the difficulty of their classes decreased and they received additional support. Controlling for classroom average skill level did not explain these changes, but *relative* skill level did. In other words, the changes in passing rates could be explained by the impact of tracking. In homogeneous (tracked) classrooms, pass rates declined for previously high-achieving students who were now at the bottom of their classroom distribution and improved for previously low-achieving students who were now at the top of their classroom distribution (the "fish pond effect"). Again, the authors argued that this had to do with the increased or decreased difficulty of the classes, rather than the impact of double dose instruction.

Research Questions

Although limited, the literature discussed above indicates a few things regarding supports for low-achieving students. First, additional instructional time is beneficial for the learning of all students, but may be particularly helpful for previously low-achieving students. Second, significant increases in instructional time may be necessary to see improved achievement. Third, additional time may be particularly beneficial in mathematics. Finally, double dose classes for low-achieving students have the potential to affect the outcomes of higher-achieving students also, particularly if they are combined with tracking.

While suggested by the prior literature, few studies have adequately described the nature of double dose policies or their impact overall and on high- and low-achieving students. Studies specifically on double dose have examined only a small number of schools (e.g., Cavanagh, 2006), failed to control for pre-existing differences between students and schools (e.g., Ney, 2010), and often failed to examine how differences in the

characteristics of double dose may affect its impact on students. As shown in the CCSR study, policy decisions such as the combination of double dose with tracking can impact its outcomes with students. None of the prior studies examined how other implementation decisions, such as which students to target and which teachers to use, may play an important part in the success or failure of the policy. Additionally, none of these studies focused on mathematics in the middle school grades.

Therefore, this chapter will use the MIST multi-state dataset to examine double dose instruction and answer the following research questions: 1) What are the characteristics of double dose instruction across middle schools in four large urban districts? 2) Are double dose policies associated with differences in average school achievement as compared to schools where double dose instruction is not provided? 3) Is the presence of double dose instruction associated with increased achievement gains for all students, or only for those in double dose classes? 4) Do differences in gains vary by the different characteristics of double dose instruction (e.g., whether it is combined with tracking, which students are targeted, and the type of curriculum used)?

Data and Measures

The data used for this analysis are also from the Middle School Mathematics and the Institutional Setting of Teaching (MIST) project. As a part of the MIST project, interview data was collected from full participant teachers, principals and mathematics instructional coaches, while class information, student achievement and student demographic data were collected for all math teachers in the school, regardless of their participation in the interviews or observations. Therefore, although only about 250 unique teachers were interviewed, 419 had associated student achievement and

demographic data. There were between 10,000 and 20,000 students associated with these 419 teachers in each year.

Tracking and Student Achievement. Using the data provided by the districts, I created three sets of variables to use in this analysis. First, I created a variable at the grade-within-school level for whether that grade was "tracked" or grouped by ability. This is the same tracking variable used in Chapter II and III, and it is an indicator of whether any classes in that school were grouped by skill level. The outcome for the second, third and fourth research questions was student achievement gains, measured using the state achievement test scores z-scored to the state distribution, controlling for the prior year's z-score. I also controlled for district to mitigate any influence of the particular test used.

The data files provided by the district also included student demographic information (race, free/reduced-price lunch, limited English proficiency and special education status), which I used as control variables in the models. I also created a variable indicating the grade level of the course, which was the grade level of the majority of students in the course. I used publicly-available data on school and district websites to create school-level control variables: number of students in the school, the percent of the test-taking population that is white, and the percent of students meeting NCLB standards. Some schools and districts also reported the percent receiving free/reduced-price lunch, English Language Learner or Special Education services. However, this was not consistently reported in all districts and years, and including these variables would have reduced the sample considerably. Therefore, I controlled only for school size, percent minority and percent meeting or exceeding NCLB standards.

This sample is slightly different than the sample used in the first two chapters, as it reflects all the students in the school, rather than only students of participating and/or observed teachers. Nonetheless, the demographics of this sample were also typical of the large, urban districts they represent. As shown in Table 23, White students were in the minority in three of the four districts, and Hispanic students were the majority in districts B and C. District A was the only district with a sizable minority of students with the race "other." In this district that was largely Asian and Native American populations. The districts were also generally low income, with 59% to 87% of students receiving free or reduced-price lunch. Finally, as each student's achievement test scores were z-scored to the state distribution, the average z-scores in each district were negative, indicating that the average achievement in these districts was one-third to one-half a standard deviation below the state average. As expected, test scores also varied significantly by student demographics, so that Black, Hispanic, LEP, special education and FRL students had lower test scores than their counterparts.

Table 23:
Student Demographics of the Sample

	Overall	District A	District B	District C	District D
Black	34%	39%	32%	29%	39%
Hispanic	40%	18%	56%	68%	6.4%
White	22%	31%	10%	2%	51%
Other	3.7%	12%	2%	0.6%	4%
Free/Reduced-Price Lunch	74%	59%	63%	87%	77%
Limited English Proficiency	14%	19%	8.9%	23%	4.8%
Male	52%	49%	52%	52%	53%
Special Education	9.7%	11%	8.3%	9%	11%
6th Grade	29%	30%	27%	24%	34%
7th Grade	35%	35%	35%	39%	31%
8th Grade	36%	34%	38%	37%	35%
Current Year Math Z score	-0.50	-0.33	-0.54	-0.51	-0.54
Prior year Math Z score	-0.36	-0.24	-0.43	-0.36	-0.35
N	62,311	10,238	15,798	18,209	18,094

Double Dose Variables. The interview data, combined with course information provided by the districts, allowed me to create double dose variables for the analysis. Teachers were asked about the types of courses they teach, the ability level of those courses, as well as the grouping of students in their schools overall. Additionally, principals, assistant principals and instructional coaches were interviewed and asked about the organization of math teaching in their school: the courses offered, how students and teachers were assigned to courses, and whether and how students are grouped by ability.

Using this interview data I first created a variable indicating whether double dose was present at that grade level in that school. For this variable, I defined double dose as a

full period of additional instruction in mathematics provided only to a subgroup of lower-achieving students. Additional characteristics will be addressed as separate variables. Teachers and principals were asked whether double dose existed in the schools. From these responses I created a binary variable that distinguishes between grades within schools where *any students* receive any form of double dose instruction and those where *no students* receive double dose instruction. I found that within schools, double dose was sometimes offered only to one or two grade levels and not all three. Therefore, this variable was applied at the grade-within-school level: all students in the same grade in the same school have the same value, but students at a different grade level in the same school may have a different value. Throughout the paper I will use "schools with double dose" or "double dose schools" as shorthand for these grades-within-schools that have double dose policies.

Occasionally participants offered conflicting reports of the presence of double dose in their school or grade level. In this case, I marked it as double dose if the principal said it existed, or if any teacher claimed to teach double dose courses. I argue that the principal has a greater knowledge of the staffing and schedule of the school than teachers, but that the existence of any double dose teachers would indicate that double dose classes must exist. Also, there were a few cases where no teachers or administrators were asked whether double dose existed in their school in a particular year. In those cases, I left the "double dose exists" variable blank, so that they would not be used as either a double dose school or as a comparison (non-double-dose) school.

Using student- and course-level files provided by the districts, I was also able to ascertain which students appeared in more than one class, and then using interview data

on whether double dose existed and who was targeted, I determined whether the second class was a true "double dose" course (provided only to a subset of lower-achieving students for a full period), rather than a tutoring session, block scheduling or other additional instruction. Students who were in an identifiable double dose course were marked as "double dose students."

Using the interview data, I created three additional variables categorizing the characteristics of the double dose policy. These were also applied at the grade-within-school level, as the double dose courses in one grade level could have different characteristics than the courses in another grade level in the same school. The first variable was a categorical variable for the target population of the double dose instruction. Teachers and principals were asked about which students were targeted for double dose instruction as a follow-up to the first question, about the presence of double dose in the school. Emerging from the data were three categories of students who could be targeted: first, what were called "Red" students, or those far below the cutoff for meeting No Child Left Behind (NCLB) standards. Second, "Yellow" or "Bubble" students, who were *near* that cutoff but did not pass[13]. Finally, some schools targeted all students below the NCLB cut score. This combines both Yellow and Red students into one group. In a few cases, schools chose instead to target students identified by the teacher, or those who were low in other subjects such as reading. I combined these

[13] "Bubble" students can also refer to students who are on the border between Proficient and Distinguished (the top two categories), but they are not examined here as a target population, because additional instruction they receive would not qualify as "double dose" in my definition.

responses into an "Other" category. Schools where no participant was asked about the target population of double dose classes were marked as missing for this variable.

The second variable I created addressed the double dose curriculum. A preliminary examination of the interviews showed that many double dose teachers said they were provided with no curriculum or other materials, and often very little direction from the principal on the focus of the class. Double dose teachers expressed that this lack of curriculum meant that they did not know the purpose of their own class, or what the students needed to learn. Therefore, I first distinguished between whether there was an adopted curriculum for the double dose classes. Then, I created categories for the curriculum being used, when one was present. As will be shown below, the most common curricula were CMP2 (the same curriculum used in the regular math classes) and SuccessMaker, a computer tutorial that emphasizes basic skills and is widely used with struggling students. Again, in cases where no participants shared this information, this variable was marked as blank. A lack of information from participants was not interpreted as a lack of curriculum for the courses.

Finally, I created an indicator for whether the same teacher taught both double dose and regular mathematics classes. On the one hand, having the same teacher in both courses might lead to additional alignment between the courses, leading to increased achievement. On the other hand, having a different teacher might provide the students with a new perspective on the material, leading to increased achievement. I attempted to create this variable from the interview data, but when I compared this variable to the course-level data received from districts, in which students were matched to their teacher, I found that they were far from aligned. In nearly half the cases, when participants said

that the same teacher was used in both classes, I could find no overlap in students, and when participants said there were different double dose teachers, I found significant overlap. Because there was also often conflict between participants in their responses to this question, I chose to use the course-level data from the districts. For the purposes of this analysis, I created a class-level variable indicating whether the majority (>50%) of the double dose students in the course had the same teacher in their second course. In a sensitivity analysis, I also examined other cut offs for this variable.

Methods

Using these variables, the first research question examined the frequencies of each of the characteristics of double dose discussed above: the combination with tracking, the target population, the curriculum used and whether the same teacher taught both double dose and regular math courses. I am interested in the extent to which "double dose" policies are similar or different across schools and districts, so I look largely descriptively, but supplement with t-tests to examine whether the variables differ significantly by district. To answer the second research question, I used data at the grade-within-school level. Each grade level within a school was marked as offering double dose ($DD_{exists}=1$) or not having double dose ($DD_{exists}=0$) in each year. Average achievement for that grade level within the school was used as an outcome. I chose to use average achievement rather than student-level achievement for two reasons. First, I am interested in how the adoption of a policy affects achievement overall in the school. The third research question will address how that policy affects individual students, but this question will examine whether the presence of double dose instruction in a grade can raise the average achievement of students in the school, regardless of whether they

actually receive the additional instructional time. Second, using a student-level achievement variable causes the problem of correlated error terms, as the independent variable is at the grade-within-school level. As using student level data would not impart any additional information to answer this question, and would cause a methodological problem, the independent variable was at the grade-within-school level (all students in this grade within the same school will have the same value of this variable), and I used average student achievement at the grade-within-school level as the outcome, as shown in Equation (11):

(11)

$$Y_{ach,s} = \beta_0 + \beta_1 S_s + \beta_2 M_{t-1,s} + \beta_3 DD_{exists,s} + e_s$$

$Y_{ach,s}$ is the average achievement of students in the grade-within-school *s*. DD_{exists} is a variable indicating whether the grade level in the school has double dose instruction of any form. I examined the β_3 coefficient to determine if the average achievement is significantly different in grades-within-schools offering double dose instruction than in grades that do not offer this instruction. This model controls for school variables to account for the possible non-random adoption of double dose policies in schools. S is a vector of school controls (size and percent minority), and $M_{t-1,s}$ is the average achievement of students in the grade-within-school from the prior year. These variables are included to account for the possibility that larger, lower-achieving and higher proportion minority schools were more likely to adopt double dose polices and also more likely to have low achievement. To account for other unmeasured differences between schools that might be correlated both with their adoption of a double dose policy and their average achievement, I also tested school fixed effects and a difference-in-

differences model in District C. These approaches are discussed in the Sensitivity Tests section.

To answer the third research question, I used a combination of the school- and student-level double dose variables as the independent variables. As *Figure 20* displays, there are three groups of students delineated by the double dose variables described above. First, there are students in grades and schools where there was no double dose present (the blue segment). Second, there are students who were in double dose schools, but were not themselves receiving double dose instruction (the green segment). These are the higher-achieving students in the double dose schools. Finally, there were students who actually received double dose instruction (the orange segment).

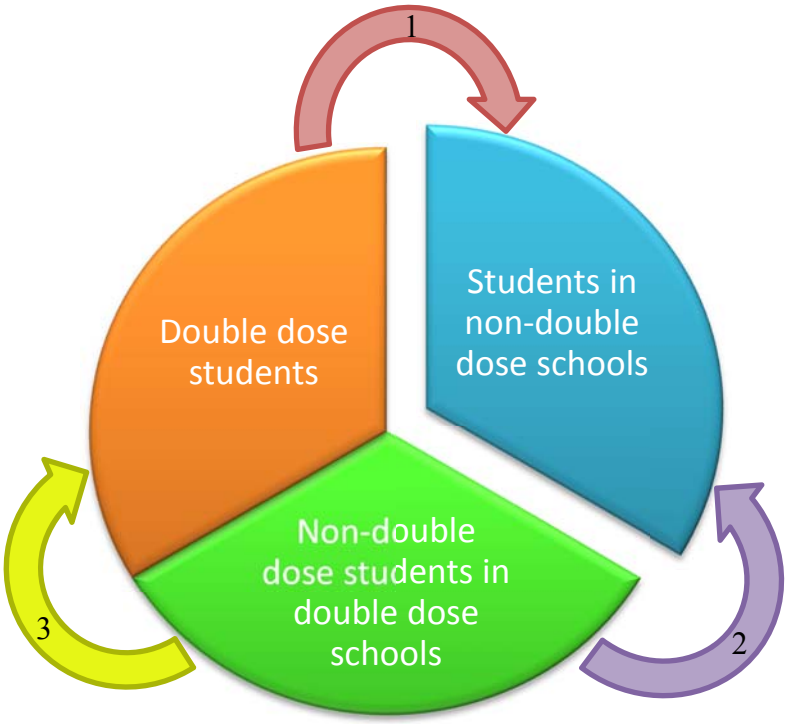

Figure 20:
Student-Level Comparisons for Research Question 3

For this research question I am interested in three comparisons, as shown by the arrows in *Figure 20*. The red arrow indicates the "double dose student effect", or the outcomes of students receiving double dose as compared to students who were not in double dose schools. The purple arrow indicates the "double dose diffusion effect," or the outcomes of higher achieving students who were in double dose schools as compared to students who were not in double dose schools. The yellow arrow shows the comparison between high- and low-achieving students in double dose schools. For the

first two outcomes, the comparison group is the same: students who were not in double dose schools. If double dose policies affected the achievement of students who were in the courses, the models testing the red arrow would be significant. If they also affected students who were in those schools, but not in those courses, the models testing the purple arrow would be significant. One reason why higher-achieving students who were in double dose courses may be affected by the policy is if their courses could move more quickly when low-achieving students were being supported. On the other hand, they may be negatively affected if the diversion of resources necessary to provide double dose harms their achievement. The third comparison, shown with the yellow arrow, would be statistically significant if the impact of double dose policies was stronger for one group (higher- or lower-achieving students) than for the other. If the comparison is not statistically significant, it means that the "double dose student effect" and the "diffusion effect" were the same size.

To test the comparisons shown in *Figure 20*, I used two student-level variables: one for double dose students ($DD_{student}$) and one for non-double dose students in double dose schools ($DD_{diffusion}$). The comparison group, then, was students who are not in double dose schools:

(12)

$$Y_{ach,i} = \beta_0 + \beta_1 X_i + \beta_2 M_{t-1,i} + \beta_3 S + \beta_4 DD_{diffusion,i} + \beta_5 DD_{student,i} + e_i$$

In this model, $Y_{ach,i}$ is the student-level achievement test score, the β_4 coefficient tests the "diffusion effect", and the β_5 coefficient is the "double dose student effect." By comparing the β_4 and the β_5 coefficients using a Wald test, I determined whether being a double dose student in a double dose school was associated with greater achievement gains than being in a double dose school but not receiving double dose instruction. If the

double dose student effect was small enough, or the diffusion effect was large enough, this difference would not be statistically significant, indicating that double dose policies had an equal impact on students in the classes and those who just attend the same schools. This model includes student-level covariates to account for the non-random sorting of students both into double-dose schools and classes. The vector X_i is student race, FRL and special education status, and $M_{t-1,i}$ is student-level prior achievement[14]. These variables have been shown to be correlated with student achievement, and if they are also correlated with the students' likelihood of being in a double dose school or class, omitting them from analysis could lead me to attribute outcomes to the impact of double dose, when they are actually due to pre-existing differences between students. Unfortunately, because I do not have longitudinal data at the student level, I could not examine student fixed effects. This is a limitation of this analysis that is discussed further in the Limitations section. As in the previous model, S is a vector of school-level covariates (size, percent minority and average prior achievement).

Finally, to answer the fourth research question, I broke down the double dose student and diffusion effect variables by the types of double dose instruction found in my investigation of the first research question, as well as by the student's prior achievement. Because this is an extension of the third research question, I am still interested in the comparisons shown in *Figure 20*. Therefore, I interacted both the $DD_{diffusion}$ and $DD_{student}$ variables with the characteristics of double dose. First, I found that some schools

[14] Although we requested district data on the achievement of students from two years prior, this variable was missing on more than 50% of students. Imputing a value when it is missing at this frequency is problematic, so I excluded it from the analysis.

combine double dose instruction with tracking, while others use double dose instruction in untracked grades. Therefore, I added an interaction between tracking and the diffusion and student effect variables in Equation (12) to create Equation (13):

(13)

$$Y_{ach,i} = \beta_0 + \beta_1 X_i + \beta_2 M_{t-1,i} + \beta_3 S + \beta_4 DD_{diff,tracked} + \beta_5 DD_{std,tracked} + \beta_6 DD_{diff,untracked} + \beta_7 DD_{std,untracked} + e_i$$

$DD_{diff,tracked}$ is the diffusion effect for tracked students, $DD_{diff,untracked}$ is the diffusion effect for untracked students, $DD_{std,tracked}$ is the double dose student effect for tracked students, and $DD_{std,untracked}$ is the double dose student effect for untracked students. This model continues to control for student- and school-level control variables. This answers two questions: first, is the "double dose student effect" (comparison 1 in *Figure 20*) larger or smaller if double dose is combined with tracking? To answer this I used a Wald test to compare the β_5 and β_7 coefficients. Second, is the "double dose diffusion effect" (comparison 2 in *Figure 20*) larger or smaller if double dose is combined with tracking? To answer this I used a Wald test to compare the β_4 and β_6 coefficients.

Next, I interacted the $DD_{diffusion}$ and $DD_{student}$ variables with the indicators on the target population of the double dose classes. I divided the possible target groups into four categories: "Red," or the lowest-achieving students, "Yellow," or students close to passing, but still below standards, "Below Cut Score," and "Other." Each category was interacted with the double dose variables as shown in Equation (14):

(14)

$$Y_{ach,i} = \beta_0 + \beta_1 X_i + \beta_2 M_{t-1,i} + \beta_3 S + \beta_4 DD_{diff,red} + \beta_5 DD_{std,red} + \beta_6 DD_{diff,yellow} + \beta_7 DD_{std,yellow} + \beta_8 DD_{diff,cut} + \beta_9 DD_{std,cut} + \beta_{10} DD_{diff,other} + \beta_{11} DD_{std,other} + e_i$$

Here $DD_{diff,red}$ is the diffusion effect when the lowest ("red") students are targeted, $DD_{std,red}$ is the double dose student effect when the lowest students are targeted, and so on. I compared the coefficients for Red, Yellow and Cut Score diffusion effects and the Red, Yellow and Cut Score double dose student effects using Wald tests. Although the "Other" category was included, it combines a wide variety of policies across only a few schools, so I did not compare its coefficients with the policies using test scores to target students.

Third, I examined interactions between the student's own prior achievement and the double dose variables, as shown in Equation (15). The literature discussed above suggested that double dose instruction may be more beneficial to previously low-achieving students. While the previous interaction examined the target group of students, this interaction examined the actual prior achievement of the students receiving double dose instruction. It is possible that, regardless of who is targeted, the lowest achieving students still benefit the most from double dose instruction. I examined prior achievement first as a continuous variable and then as categorical, to account for the possibility that the relationship is non-linear.

(15)
$$Y_{ach,i} = \beta_0 + \beta_1 X_i + \beta_2 M_{t-1,i} + \beta_3 S + \beta_4 DD_{diffusion} + \beta_5 DD_{student} + \beta_6 DD_{diff} * M_{t-1,i} + \beta_7 DD_{std} * M_{t-1,i} + e_i$$

Fourth, I added interactions between the double dose variables and the choice of curriculum. When a curriculum was present, it was overwhelmingly either CMP2 or SuccessMaker, but there was a small group of schools that used another curriculum. I

divided curriculum type into these four categories: CMP2, SuccessMaker, other curriculum, and no curriculum. Equation (16) shows the addition of these interactions:

(16)

$$Y_{ach,i} = \beta_0 + \beta_1 X_i + \beta_2 M_{t-1,i} + \beta_3 S + \beta_4 DD_{diff,cmp} + \beta_5 DD_{std,cmp} + \beta_6 DD_{diff,sm} + \beta_7 DD_{std,sm} + \beta_6 DD_{diff,none} + \beta_7 DD_{std,none} + e_i$$

The $DD_{diff,cmp}$ is the diffusion effect when the adopted curriculum was CMP2, the $DD_{diff,sm}$ is the diffusion effect when the adopted curriculum was SuccessMaker, and $DD_{diff,none}$ is the diffusion effect when there was no curriculum. As with the target group of students, the "other" category in curriculum was small and diverse in the type of materials used, so the interactions were included only so that the comparison group here is still students who are in grades-within-schools where there is no double dose.

Finally, I examined an interaction between the double dose student and diffusion effects and whether the same teacher was used for double dose and regular mathematics classes:

(17)

$$Y_{ach,i} = \beta_0 + \beta_1 X_i + \beta_2 M_{t-1,i} + \beta_3 S + \beta_4 DD_{diff,same\ t} + \beta_5 DD_{std,same\ t} + \beta_6 DD_{diff,other\ t} + \beta_7 DD_{std,other\ t} + e_i$$

Here $DD_{diff,same\ t}$ is the diffusion effect when the same teacher was used in both double dose and regular math classes, and $DD_{diff,other\ t}$ was the diffusion effect when a different teacher was used for double dose courses. I compared the β_4 and β_6 coefficients using a Wald test to examine whether the diffusion effect was different under these different circumstances, and compared the β_5 and β_7 coefficients to test whether the student effect was different.

Descriptive Statistics on the Sample

Sixty-one percent of students were in grades in schools where double dose instruction was provided, but only 13% of students in double dose schools could be identified as receiving double dose instruction. This reinforces the concept of double dose as a program applied to a minority of students. The demographics of double dose courses were significantly different from those of regular courses. These courses had a significantly lower proportion of Black and a higher proportion of Hispanic students than regular math courses. The concentration of students receiving free or reduced-price lunch or special education services was higher in double dose courses, but there was no difference in the proportion of students identified as Limited English Proficient (LEP). Finally, as shown in *Figure 21*, both prior and current achievement of students in double dose classes was significantly lower than those who were not in double dose courses.

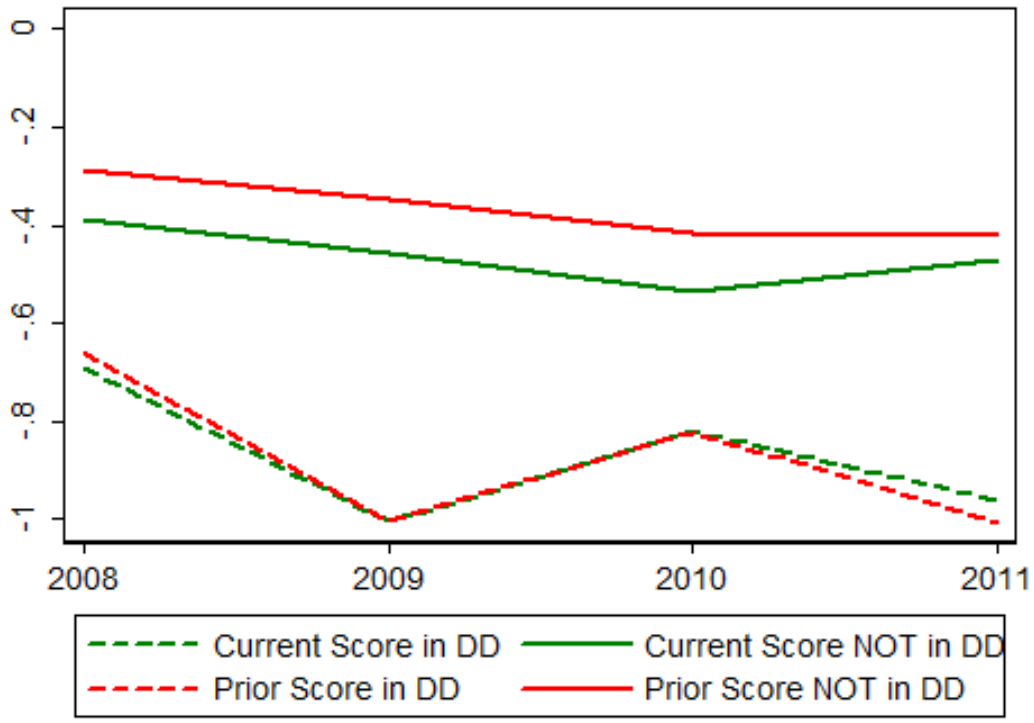

Figure 21:
Line Graphs of Unadjusted Current and Prior Student Achievement over Time by Participation in Double Dose Classes

School and teacher characteristics also varied by whether double dose was offered at the school. Schools with double dose in any grade level were significantly smaller, had lower proportions of minority students, and lower average prior achievement (see Table 24). The values of these variables were taken from publicly-available data on school and district websites, and so do not reflect any possible bias from non-random selection of teachers due to refusal to participate. On the other hand, teacher variables were determined using the MIST teacher survey, so I can only examine the teachers in the MIST sample. The teachers in our study who taught in grades and schools with double dose had significantly more experience (12.2 compared to 8.8 average years of

experience), and the proportion of white teachers was significantly higher. Double dose was spread fairly evenly across grades, with about 31% of double dose classes in 6th grade, 34% in 7th grade and 35% in 8th grade.

Table 24:
School Characteristics of Double Dose and Non-Double Dose Schools

	Double Dose	Non-Double Dose
School Characteristics		
Percent Free/Reduced-Price Lunch	75.1%	74.8%
Percent Limited English Proficient	19.9%	15.4%
Percent Special Education	14.0%	11.7%
Percent Minority	79.3%	86.4%
Average School Size (# of students)	631	796
Percent Meet/Exceed Standards	43.0%	59.1%
Average prior year z-score	-0.531	-0.328
Teacher Characteristics		
Average years of experience	12.2	8.8
Percent fully certified	96.3%	93.1%
Percent white	78.6%	52.4%
Average age	43	41
Number of methods courses taken	3.5	3.2
Number of math content courses taken	3.0	2.8
Number of advanced math courses taken	2.4	2.3

The association between demographics and both the double dose variables and student achievement indicate that the models used to answer my research questions must control for student race, free/reduced-price lunch and Special Education status as well as school size, proportion minority and percent meeting or exceeding standards at the school level. Unfortunately, I cannot control for teacher characteristics, because I only have data on the teachers who were in our study, so excluding missing values would severely

reduce the sample, including eliminating many students receiving double dose instruction.

Results

Research Question 1: What are the characteristics of double dose instruction across middle schools in four large urban districts? About 63% of schools in the MIST study had double dose classes, and this proportion increased from 51% in 2008 to 70% in 2010, and then leveled off at 69% in 2011. The prevalence of double dose also varied significantly by district. As shown in Figure 22, all schools in district D had double dose in all four years, and none in district C had double dose before year 3, which indicates that most of the variation necessary for analysis in question two will come from districts A and B.

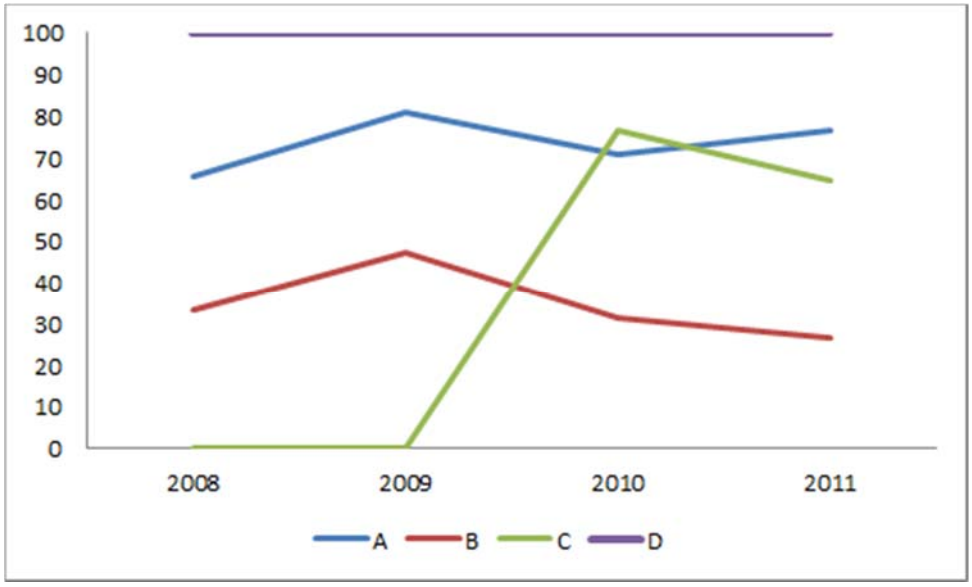

Figure 22:
Line Graphs of the Percent of Grade Levels in Schools with One or More Double Dose Classes over Time

Across districts, there was a fairly even split between schools that combined double dose with tracking and those that did not. About 37% of grade levels had no double dose classes, 38% had double dose and tracking combined, and 25% had double dose classes and no tracking (only one course level in mathematics at that grade level). As expected, this varied significantly by district as well, so that in district B most schools that had double dose also had tracking, and in district D, where all schools had double dose, most schools did not have tracking.

Double dose courses in these schools differed in the populations they targeted. About 58% targeted all students below the proficiency cutoff, including both marginal and very low-achieving students. Twenty percent of schools said they targeted only the lowest-achieving or "red" students, and 12% focused on students who were close to passing ("yellow" students). Another 8% selected students for double dose instruction based on reading scores or other criteria besides math test scores, such as teacher recommendations. In 2% of grade levels, there was not enough information in the interviews to determine the target population of double dose instruction. Again, this varied significantly by district. Schools in District C and D were most likely to target all students below a cut score, while those in District A and District B were more likely to target Red or Yellow students.

Nearly half (43%) of grades with double dose courses did not have any adopted curriculum in place, while 27% used SuccessMaker, an individualized computer-adaptive software (Pearson, 2012). Another 8.7% used the Connected Math Project (CMP2) curriculum that was adopted by the district as the main curriculum, 12.6% used an

unspecified other curriculum, and 7.7% did not have enough information to determine. In District D, nearly 70% of double dose schools used SuccessMaker, while in District B, 85% did not have an adopted curriculum. This variation by district was statistically significant ($p<0.01$).

As shown in Table 25, 29.7% of double dose courses had *all* the same students as those in another regular math course, another 37.3% of double dose courses pulled some students from the same teacher, and other students from a different teacher, and 33.1% had a different teacher for double dose than for regular math in all cases. This variation stems in part from the combination of tracking and double dose. In 54% of the cases where all students in a double dose and a regular course were taught by the same teacher, double dose was combined with tracking. So, low-achieving students took their regular math course together and an additional math course together, separated from higher-achieving students at all times. This also varied by district. In District A, more than 42% of the double dose classes had the same teacher as in the regular course for all students. In District B this was never the case.

Table 25:
Did double dose courses have the same teacher as regular math instruction courses?

	Overall
Same teacher for all students in the course	29.7%
Same teacher for some students in the course	37.3%
Different teacher for all students in the course	33.1%

Overall, the prevalence of double dose increased across the four years, but the characteristics of what was called double dose varied greatly. It was combined with

tracking about half the time and was often extended to all students below a cut score on state tests, but was targeted to a smaller group nearly a third of the time. Nearly half of all double dose classes did not have an adopted curriculum, and 33% used entirely different teachers for double dose and regular instruction classes. Each of these characteristics varied by district and may have its own relationship with student achievement, so this variation in policy implementation characteristics will be important to address in Research Question 4.

Research Question 2: Are double dose policies associated with differences in average school achievement as compared to schools where double dose instruction is not provided? As shown in Table 26, the existence of double dose instruction in schools was associated with lower achievement on average. This is not surprising, as low-achieving schools may be more likely to implement double dose policies. However, controlling for average prior achievement, as well as school characteristics, grades-within-schools with double dose still had average achievement $1/10^{th}$ of a standard deviation lower than grades-within-schools without double dose. This indicates that, among schools similar in prior achievement and other characteristics, the presence of double dose policies in schools is associated with *lower* achievement than the absence of those policies. One reason for this, which cannot be tested here, may be the diversion of resources from other supports for students. On the other hand, it may be due to unobserved differences between the schools or grade levels that choose to adopt double dose policies and those that do not. To test this, I examined school-level fixed effects, which look at the relationship between double dose policy and average achievement *within schools*. This is possible because some schools either adopted or abandoned

double dose policies over time, so this model examined how average achievement changed in a school when double dose policies changed. In this model I also controlled for district, study year, grade level, and student demographics. The introduction of school fixed effects did not change the size or the significance of the coefficient. This indicates that the negative relationship between the existence of double dose and grade-level average achievement could not be accounted for by pre-existing unobserved differences between schools. I test this further using a difference-in-differences model in district C in the Sensitivity Tests section.

Table 26:
Models Predicting the Relationship between Double Dose Policies and School Average Achievement

	No Controls		With Controls	
Double Dose Exists	-0.29***	(0.06)	-0.08**	(0.03)
Year 3	-0.05	(0.06)	-0.06	(0.03)
Year 4	-0.06	(0.07)	-0.08*	(0.03)
District 2	-0.09	(0.06)	-0.15***	(0.03)
District 4	-0.27***	(0.07)	0.01	(0.05)
Avg Prior achievement			0.71***	(0.04)
School size			-0.00**	(0.00)
School % minority			-0.12	(0.08)
School % meet/exceed			0.71***	(0.09)
Constant	-0.35***	(0.04)	-0.22**	(0.08)
Observations	328		322	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Research Question 3: Do these differences in gains vary by whether the student was in a double dose class? For this research question, I moved to student-level data, to examine the impact of double dose on individual students' achievement. These models examine 1) the double dose student effect (comparing the achievement of double dose students to the achievement of students in non-double dose schools), 2) the double dose diffusion effect (comparing the achievement of non-double dose students in double dose schools to the achievement of students in non-double dose schools, and 3) the relative size of the two effects (comparing double dose students to non-double dose students in the same schools). As shown in Table 27, both double dose and non-double dose students in double dose schools had significantly lower achievement than those in non-double dose schools. When controlling for prior achievement and demographic characteristics, the "double dose student effect" is reduced, but does not disappear, and

the "diffusion effect" is reduced only slightly. With covariates included, the difference in the size of these double dose student and diffusion effects was not statistically significant. So, double dose policies were associated with *lower* achievement gains for all students in double dose schools, and there was no difference in the effect for students who were in double dose classes and those who were not. One reason why high-achieving students in double dose schools might have lower achievement than students in non-double-dose schools is the diversion of resources. Double dose policies are expensive, and the introduction of such a policy is likely to mean the reduction of other services or supplies. It is possible that double dose programs cause negative effects for those not in the classes by reducing the supports they receive. This would also be associated with lower achievement for the double dose students if any benefits of additional instructional time were not enough to overcome the detrimental impact of the reduction of other resources or of the labeling effect of assigning students to these courses. Unfortunately, I could not determine from the MIST data whether either of these causes explains the negative double dose effects.

Table 27:
Models Predicting the Achievement of Double Dose and Non-Double Dose Students Compared to Students in Schools without Double Dose

	No Controls		Add Demographics	
Double dose student effect	-0.48***	(0.02)	-0.04***	(0.01)
"Diffusion Effect"	-0.06***	(0.01)	-0.04***	(0.01)
Prior achievement			0.76***	(0.00)
Black			-0.12***	(0.01)
Hispanic			-0.03***	(0.01)
Other			0.02	(0.01)
FRL			-0.04***	(0.01)
SPED			-0.18***	(0.01)
School Size			-0.00**	(0.00)
School % minority			0.12***	(0.02)
Year 2			-0.06***	(0.01)
Year 3			-0.08***	(0.01)
Year 4			-0.03***	(0.01)
District 2			-0.22***	(0.01)
District 3			-0.21***	(0.01)
District 4			-0.02	(0.01)
7th grade			0.05***	(0.01)
8th grade			0.12***	(0.01)
School prior % meet/exceed			0.58***	(0.03)
Constant	-0.43***	(0.01)	-0.28***	(0.02)
Difference in Effects	0.40 (p<0.001)		0.00 (p=0.86)	
Observations	57,097		49,423	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Another potential explanation for these effects is pre-existing unobserved differences between schools, a possibility I examined using school fixed effects. When school fixed effects were added, *neither* the double dose student effect nor the diffusion effect was statistically significant. This indicates that the apparent relationship between double dose policies and lower student achievement may be due to unmeasured school factors associated with both the existence of a double dose policy and the potential for

lower achievement gains. One such difference may be the different characteristics of double dose instruction.

Research Question 4: Do differences in gains vary by the different characteristics of double dose instruction? While the direction of the results found in Research Question 3 was the opposite of expected, it is important to examine whether different characteristics of double dose might be associated with different outcomes. As shown in Research Question 1, there was a significant amount of variation in the characteristics of double dose as it was implemented in schools. I focused on five such characteristics in this section: 1) the combination of double dose and tracking, 2) the population targeted by double dose classes, 3) the students' prior achievement, 4) the curriculum used in double dose classes, and 5) whether the double dose and regular classes were taught by the same teacher.

Double Dose and Tracking. First, some schools that implemented a double dose policy for low-achieving students also separated students into course levels (e.g., "honors" and "regular") in their main math class. As shown in Table 28, controlling for prior achievement and demographics, the double dose student effect was negative in both tracked and untracked schools, and there was no difference in the size of the effect between settings, when examined using a Wald test comparing the coefficients. The diffusion effect was negative and statistically significant in both tracked and untracked schools, but significantly larger in untracked schools ($p<0.001$). I examined this question using school fixed effects, and found that while there was no significant within-school student or diffusion effect in tracked schools, the diffusion effect remained significant in untracked schools. So, within untracked schools, the higher-achieving students who were

in grade levels that had double dose offered score about 0.06 standard deviations below students who are in grade levels with no double dose, controlling for other factors. The negative "diffusion effect" seems to be found largely in untracked schools, where students who were in double dose and those who were not in double dose share their regular math class.

Table 28:
Models Predicting the Interaction between Double Dose Effects and Whether the School is Tracked in Mathematics

	Estimate	Standard Error
Double Dose student Effect		
Tracked	-0.06***	(0.01)
Untracked	-0.04*	(0.02)
Diffusion Effect		
Tracked	-0.03***	(0.01)
Untracked	-0.08***	(0.01)
Prior achievement	0.76***	(0.00)
Black	-0.12***	(0.01)
Hispanic	-0.03***	(0.01)
Other	0.02	(0.01)
FRL	-0.04***	(0.01)
SPED	-0.18***	(0.01)
School Size	-0.00***	(0.00)
School % minority	0.13***	(0.02)
Schl % meet/exceed prior	0.56***	(0.02)
Year 2	-0.06***	(0.01)
Year 3	-0.08***	(0.01)
Year 4	-0.03***	(0.01)
District 2	-0.22***	(0.01)
District 3	-0.21***	(0.01)
District 4	-0.01	(0.01)
7th grade	0.05***	(0.01)
8th grade	0.12***	(0.01)
Constant	-0.26***	(0.02)
Student Effect Difference	0.02	
Diffusion Effect Difference	0.05***	
Observations	49,423	

Standard errors in parentheses, + p<0.10 * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Target Population of Double Dose Classes. Second, I examined whether the relationship between double dose and achievement varied by who was targeted by the double dose classes. I divided the target into three categories of interest: "Red," or the lowest-achieving students, "Yellow," or students close to passing, but still below standards, and "Below Standards." As shown in Table 29, the double dose student effect (the relationship between achievement and participation in a double dose course) was statistically significant and negative when all students below a cut score or students far below the cut score were targeted, but it was not significantly different than zero when students near the cut score were targeted. However, when comparing these coefficients using Wald tests, there was no statistically significant difference in their size.

Table 29 also shows that the diffusion effect (the relationship between achievement and being in a double dose school and *not* receiving double dose instruction) was statistically significant and negative for all three possible target groups. When comparing these coefficients, the size of the effect was significantly more negative when Yellow students were targeted than when either Red or all students below the cut score were targeted.

Table 29:
Models Predicting the Interaction between Double Dose Effects and the Group of Students Targeted for Double Dose Instruction

	Estimate	Standard Error
Double Dose student Effect		
Red students are targeted	-0.04*	(0.02)
Yellow students are targeted	-0.02	(0.04)
Students below cut score are targeted	-0.05**	(0.02)
Diffusion Effect		
Red students are targeted	-0.05***	(0.01)
Yellow students are targeted	-0.08***	(0.01)
Students below cut score are targeted	-0.03***	(0.01)
Observations	49,048	

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
These models control for the same demographic variables as were used above

These findings indicate that the target group of students may have a different impact on students taking those courses than on the students in the same schools and classrooms who were not in double dose. While there was no statistically significant difference in the double dose student effect depending on who was targeted, the diffusion effect was significantly worse if only students near the cut score were targeted. This may be because resources were diverted from regular instruction, but the number of students benefiting from double dose was not large enough to have a positive impact on the classrooms they share with non-double-dose students.

Prior Achievement. In addition to the target population of double dose instruction, the student's own prior achievement may affect the impact of double dose. To test this, I examined an interaction between prior achievement and both the student

and the diffusion effect. Neither interaction was statistically significant, which implies that double dose courses affected students at different incoming achievement levels about the same. However, it is possible that the relationship is non-linear. Therefore, I also examined this same question with prior achievement as a categorical variable, as shown in Table 30. This compared students in each of six categories of prior achievement to students in non-double-dose schools at the same level of prior achievement, and allowed the relationship to be different at each level, rather than requiring it to increase linearly from one level to the next. The effects remained negative, but their statistical significance and size varied by category.

Table 30:
Models Predicting the Interaction between Double Dose Effects and Prior Achievement Categories

	Range of Prior Achievement Z Scores					
	<=-2	-2 to -1	-1 to 0	0 to 1	1 to 2	>2
Double Dose student effect	-0.090*	-0.022	-0.047*	-0.062	-0.155*	-0.337***
	(0.043)	(0.022)	(0.019)	(0.035)	(0.076)	(0.046)
Diffusion Effect	-0.038	0.003	-0.042***	-0.060***	-0.047	-0.107***
	(0.033)	(0.015)	(0.012)	(0.014)	(0.027)	(0.027)
Observations	2641	10965	18945	12441	3717	7952

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
These models control for the same demographic variables as were used above

As *Figure 23* shows, the double dose student effect was most negative among students whose prior achievement was more than two standard deviations above average, but it was also significantly negative for students whose prior achievement was two or more standard deviations *below* average. The bowed shape of this predicted distribution

suggests that double dose classes were associated with worse outcomes for those at the lowest and highest ends of the distribution, when compared to similar students in non-double dose schools. The diffusion effect was also most negative among previously high-achieving students, but it was also statistically significant and negative among students near the state average. The diffusion effect was not significantly different than zero for very low-achieving students. The finding that the diffusion effect was most negative when "Yellow" students are targeted and least negative for double dose students near the center of the distribution presents a policy dilemma. This, combined with the fact that the "diffusion effect" was negative for students near the center of the distribution indicates that singling out some students for this instruction while excluding others may have a harmful effect on those who are left out. This may be due to the diversion of resources from tutoring or other additional supports for non-double-dose students.

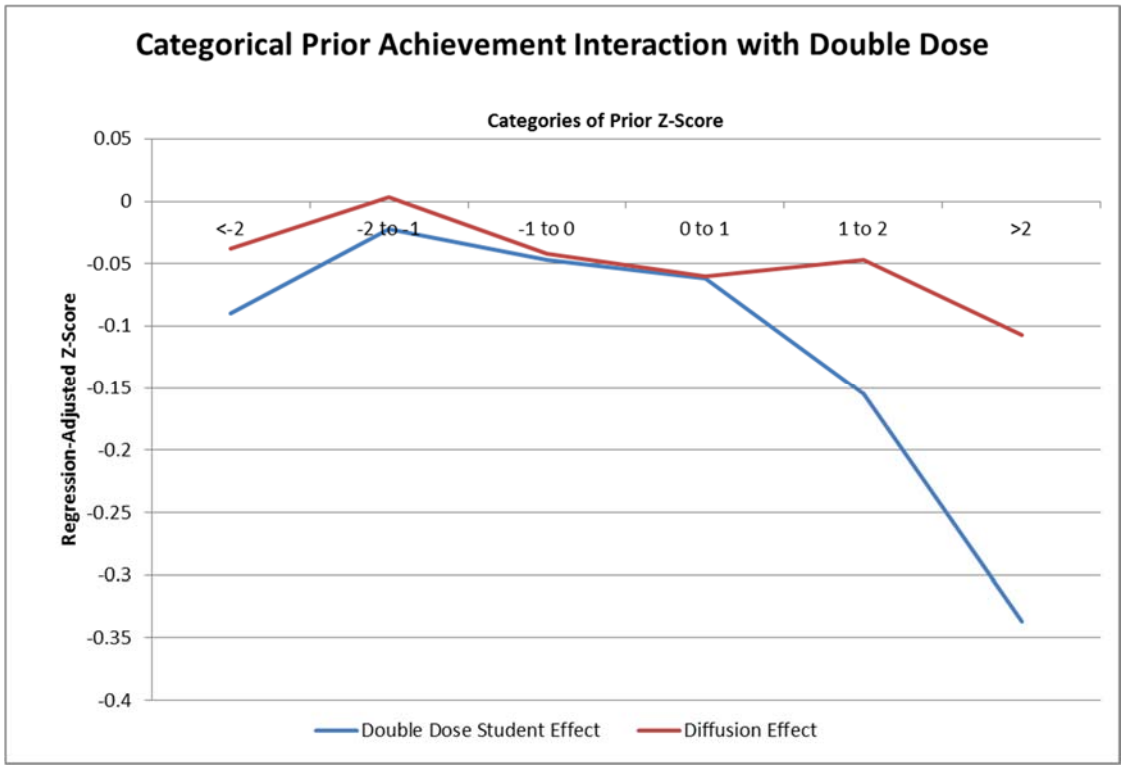

Figure 23:
Line Graphs of Regression-Adjusted Z-Scores for Double Dose Student and Diffusion Effects by Categories of Prior Achievement

Curriculum Used in Double Dose Classes. Third, I tested whether the presence and type of curriculum was related to either the double dose student effect or the diffusion effect on non-double dose students in double dose schools. The comparison group for this model was students who were not in double dose grades or schools. As shown in Table 31, both the student and the diffusion effects were statistically significant and negative in settings where there was a no adopted curriculum for double dose classes and where CMP2 was the adopted curriculum. However, both effects were small, but positive when schools used SuccessMaker in double dose courses. The difference between effects was statistically significant when using a Wald test, indicating that

schools using SuccessMaker in their double dose courses have significantly better predicted student outcomes than those using other or no curricula in their double dose courses.

Table 31:
Models Predicting the Interaction between Double Dose Effects and Type of Curriculum Used in the Double Dose Classes

	Estimate	Standard Error
Double Dose student Effect		
With CMP2	-0.08*	(0.03)
With SuccessMaker	0.05*	(0.02)
Without a curriculum	-0.06***	(0.02)
Diffusion Effect		
With CMP2	-0.09***	(0.02)
With SuccessMaker	0.03*	(0.01)
Without a curriculum	-0.04***	(0.01)
Observations	49,048	

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
These models control for the same demographic variables as were used above

Double Dose Teachers. Finally, I examined the impact of whether the double dose and regular classes were taught by same teacher. Some double dose students had the same teacher for both their regular and their double dose classes. Usually this did not mean that all students were in both classes together, but rather that some students were in both classes. For the purposes of this analysis, I created a class-level variable indicating whether most (>50%) of the double dose students in the regular math course had the same teacher in their double dose course. I found that concentrating double dose students in courses in this way was associated with significantly less negative effects on the other students in double dose schools, but made no difference for double dose students (see Table 32). As discussed in the first research question, more than half of the time schools

that used the same teachers in double dose and non-double dose classes were also tracked, so this finding and the finding that the negative diffusion effect was smaller in tracked settings may be related.

Table 32:
Models Predicting the Interaction between Double Dose Effects and whether the Same Teacher Taught Both Courses

	Estimate	Standard Error
Double Dose student Effect		
Same teacher for >50% of students	-0.04***	(0.01)
Same teacher for <50% of students	-0.06***	(0.02)
Difference in the Student Effect	0.02	
Diffusion Effect		
Same teacher for >50% of students	-0.04***	(0.01)
Same teacher for <50% of students	-0.08***	(0.01)
Difference in the Diffusion Effect	0.04***	
Observations	49,423	

+p<0.10 * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
These models control for the same demographic variables as were used above

As a sensitivity test, I examined alternative cutoffs for the percent of double dose students with the same teacher. I found similar results when I used 75 percent and when I used 100 percent as a cut off for the "same teacher" variable. Additionally, when I entered the percent of students with the same teacher as a categorical variable, I found that both the double dose student effect and the diffusion effect were significantly more negative when *either* zero to 24 percent *or* 25 to 49 percent of students shared the same teacher in double dose than when any higher percentage shared the same teacher. This indicates that the fifty percent cutoff is a valid choice.

Combining Characteristics. The fact that different characteristics were associated with different student and diffusion effects suggests that there may be a

"perfect" and an "imperfect" double dose policy, where the "perfect" policy does not have the same detrimental effects on students. However, the characteristics associated with less negative effects were different for the students in the double dose courses and the students who were in double dose schools but not in those courses. The double dose diffusion effect was less negative when the grade level was tracked or SuccessMaker was the adopted curriculum, and more negative when "Yellow" students were targeted[15]. On the other hand, the double dose student effect was positive when the SuccessMaker curriculum was used, but none of the other characteristics had a significant impact. I examined the "perfect" conditions as defined by those best for the diffusion effect and found that, as shown in Table 33, there were few students who could be identified as in double dose classes under these conditions.

Table 33:
Number of Students in "Perfect" Double Dose Conditions by Year

	Double Dose Students	Non-Double Dose Students in Double Dose Schools
Year 1	87	849
Year 2	114	1,511
Year 3	259	2,352
Year 4	158	1,986

[15] It was also less negative when the same teacher was used in double dose and regular classes, but in all cases where "Yellow" students were not targeted, SuccessMaker was the adopted curriculum, and the grade level was tracked, the same teacher was used for both classes. Colinearity, therefore, made it impossible to include this variable.

Nonetheless, I explored a model comparing the double dose student and diffusion effects in "perfect" and "imperfect" conditions. As shown in Table 34, both the double dose student and the diffusion effects were *positive* under these "perfect" conditions. This indicates that double dose instruction not targeted at "Yellow" or "Bubble" students, which used the SuccessMaker curriculum, where both classes were taught by the same teacher, and the grade level was grouped by achievement for regular math instruction was associated with higher achievement when compared to both other double dose policies and when compared to situations with no double dose. The students in such courses had predicted achievement about 0.06 standard deviations higher than similar students who were not in double dose schools or were in other forms of double dose. Students who were in double dose schools but not in double dose courses had predicted achievement about 0.09 standard deviations higher than similar students who are not in double dose schools, and 0.11 standard deviations higher than similar students in schools with a different type of double dose. Although the number of observations was small, this finding shows that certain forms of double dose may benefit both the students in the courses and their classmates not receiving double dose instruction, while other forms actually harm students in those schools. In fact, the small sample indicates an unfortunate finding: the types of double dose that may actually support student achievement growth are rarely being implemented in the schools in these districts. Only four schools ever employed double dose with these characteristics, and only one of these schools kept this form of double dose for the full four years of our study.

Table 34:
Model Predicting Double Dose Effects in "Perfect" and "Imperfect" Conditions as compared to Schools without Double Dose

	Estimate	Standard Error
Double Dose student Effect		
"Perfect" Conditions	0.061*	0.025
"Imperfect" Conditions	0.005	0.012
Difference in the Student Effect	0.055*	
Diffusion Effect		
"Perfect" Conditions	0.085***	0.010
"Imperfect" Conditions	-0.024***	0.006
Difference in the Diffusion Effect	0.109***	
Observations	49,048	

+p<0.10 * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
These models control for the same demographic variables as were used above

Including school fixed effects in the "combined characteristics" model reduced the coefficients slightly, but they remained statistically significant. This indicates that the positive relationship between "perfect" double dose conditions and "imperfect" double dose conditions cannot be entirely attributed to unobserved school factors[16].

Sensitivity Tests

Imputation for Missing values. As a sensitivity test to the results reported above, I imputed for missing values on the independent variables in each model. While the student achievement variable was missing in about 3,000 cases, these values were not possible to impute because it was used as a dependent variable. When these 3,000 cases were dropped, there were five variables created from the district dataset with missing values: prior achievement, race, free/reduced-price lunch status, special education status

[16] This comparison is possible because schools with "perfect" double dose conditions in some grades or years occasionally had "imperfect" double dose conditions in another grade or year.

and the "same teacher" variable (whether the same teacher was used in regular and double dose classes)[17]. While 7,673 students were missing their prior achievement, only about 20 cases were missing the other variables. Therefore, missing values were imputed using multivariate normal regression in Stata, which uses an iterative Markov Chain Monte Carlo (MCMC) method to fill in five plausible values based on the values in variables with no missing data (Statacorp, 2009, p. 145). Using these imputed values, I re-ran the regressions for research questions two through four, and found no differences in the significance or direction of the coefficients, and only small differences in the size.

Difference-in-Differences. As another sensitivity test, I examined a difference-in-differences model in district C, where double dose policies were instituted in the third year of the study in some schools. Unfortunately, because I do not have longitudinal data at the student level, it was only possible to apply this sensitivity test to the second research question, which used data at the grade-within-school level. The difference-in-differences model examines whether the difference between average achievement in double dose schools and non-double dose schools was significantly different before and after the policy went into effect. As expected, before the policies went into effect, the average achievement of District C schools that chose to implement double dose policies was low and declining, but at about the same rate as schools that chose not to adopt double dose (see Figure 24). Post-policy, the achievement in double dose schools continued to decline, while the achievement in non-double dose schools increased.

[17] I did not attempt to impute missing values on the characteristics of double dose that were missing because I was not confident that the other variables in the dataset could reliably predict these policy decisions.

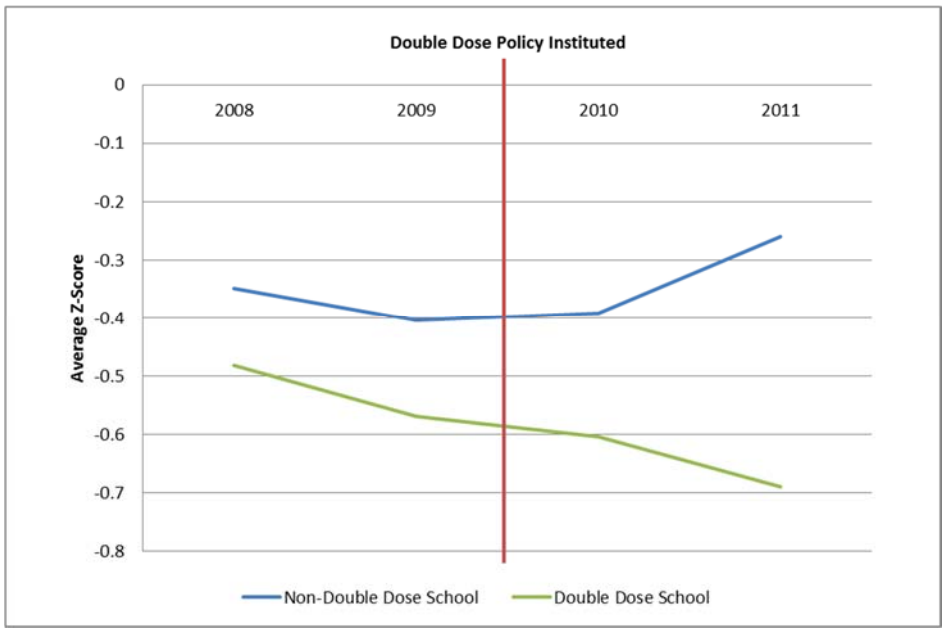

Figure 24:
Unadjusted Difference in School Achievement in District C Before and After Double Dose Policies Appeared in Schools

However, as shown in Table 35, this difference-in-differences was not statistically significant when controlling for other school-level demographic variables. This sensitivity test, combined with the school-level fixed effects and the findings in response to research question 4, provide evidence that the negative relationship between adopting double dose policies and achievement is due to school factors, many of which are policy decisions on what form of double dose to implement.

Table 35:
Difference-in-Difference Model for School-Level Impact of Double Dose Policies in District C

	No Controls		Add Demographics	
Difference in differences	**-0.15**	**(0.27)**	**-0.02**	**(0.14)**
Double dose school effect	0.16	(0.19)	0.00	(0.14)
Post-2009 effect	0.01	(0.25)	-0.09	(0.13)
Average prior achievement			0.19*	(0.07)
School size			0.00	(0.00)
School % minority			-8.79***	(1.52)
School % meet or exceed standards			0.65*	(0.28)
Constant	-0.53**	(0.17)	7.53***	(1.43)
Observations	68		62	

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Limitations

Although this analysis significantly advances our understanding of double dose instruction, its various forms and its impact, the data are correlational in nature. I do not have longitudinal data at the student level, so I cannot examine the impact of students moving from double dose to non-double dose, or vice versa. If there is non-random sorting of students into double dose classes and schools, which is not absorbed into the student's prior achievement or the demographic controls used here, then the results I found above could be attributed to pre-existing differences in students, rather than the effect of double dose policies. I also only have one district in which I can examine achievement before and after schools passed double dose policies. Although District C did begin to implement double dose in some schools in the middle of the MIST study, the lack of student-level longitudinal data made it impossible to run a regression discontinuity or difference-in-differences to examine research questions 3 or 4. By controlling for prior achievement and comparing students in the same grades and schools,

I hope to mitigate omitted variable bias as much as possible, so that any relationship I find between double dose and achievement can reasonably be believed. However, this is not enough to make a strong causal claim that double dose policies impact student achievement in a negative way. Instead, they serve as a first step in examining the relationship and will hopefully lead to more research in the area.

The fact that the initial results were the opposite of expected, that the presence of double dose policies is associated with *lower* achievement both among the students in those courses and among students in the same schools but not in double dose courses, underlines the need for further research. My finding that the association between double dose and student achievement varies based on the characteristics of double dose, such as which students are targeted and the curriculum used, shows how much more work there is to be done. "Double dose" policies vary greatly by school and district, so adopting policies of additional instructional time for low-achieving students is not the only choice that these stakeholders must make. Choices made in implementation are vital.

In addition to the variables explored here, there may be other important factors that went unmeasured. For example, the characteristics of the teachers, when they are not the same between double dose and regular courses, are likely to impact its success. Without access to survey data on all teachers, I was unable to examine this question. The quality of instruction in the double dose courses is also likely to impact their success. Although the MIST project measured instructional quality in a sample of classrooms in each school, very few of these (one or two in each year) were double dose classrooms. Additionally, other characteristics of schools may have a relationship with both double dose and student achievement, such as the other programs offered to low-achieving

students. Finally, relationships among staff and students at the school are likely to have a great impact on the success of double dose. For example, trust among teachers and between the teachers, principals and students could lead to a more effective implementation of a policy such as this.

The source of the data on the characteristics of double dose instruction also presents a limitation. Although student achievement data was available on all math students in the school, the interview data came from only a subsample of teachers. This sample was intended to be representative of the school, but participation was voluntary, so the impressions of how double dose is organized in their school may not be the full picture. For example, if, within a school, the double dose teacher was not interviewed, the question of the curriculum used in double dose courses fell to the regular math teacher, the principal or other participants. If they were mistaken about what was happening in the double dose classroom, this would introduce error into my models.

Conclusion

Schools have adopted many approaches to dealing with the variation in incoming achievement levels in middle school. These approaches often introduce increases in instructional time, which have been shown to be associated with student achievement (e.g., Bloom, 1974; Brown & Saks, 1986; Marcotte, 2007). One recent policy that increases instructional time for low-achieving students is double dose instruction. While the prevalence of double dose has increased, there is little research on its impact. The justification for these policies seems to rest on the link between time and achievement, without much concern for the different ways the policies may actually be implemented in schools. This study advances the scholarly understanding of double dose policies by 1)

examining the differences in the characteristics of double dose policies across thirty schools in four large, urban districts; 2) examining the relationship between double dose policies and average achievement across these districts; 3) separating "student effects" from "diffusion effects" when examining how double dose is associated with student-level achievement; and 4) investigating how student-level outcomes may vary based on the characteristics of double dose.

I found that under the title of "double dose" are a wide variety of policies. Schools and districts vary in which students are targeted, who teaches the courses and what curriculum is used. By defining double dose as a full period course offered only to low-achieving students, I eliminated students in tutoring and block scheduling, but combined all of the above variation into one category: "double dose exists." The existence of double dose of any kind was associated with *lower* average student achievement at the school level, even controlling for average incoming student achievement. Likewise, the association between double dose and student achievement was negative for both the students in those courses and for the students who were not in those courses, but were in double dose schools. However, these "student" and "diffusion" effects were not statistically significant when school fixed effects were added, indicating that the relationship between double dose and student achievement may be due to school-level characteristics.

One set of such characteristics was the different forms of double dose I found in the first research question. By breaking out the relationship between achievement and double dose by the characteristics of the policy, I found that the association was less negative for double dose students when the SuccessMaker curriculum was used. For the

students in double dose schools but not receiving double dose instruction, the SuccessMaker curriculum was also associated with less negative impact, as was ability grouping at the classroom level. On the other hand, the diffusion effect was more negative when Yellow students were targeted.

In an important finding, when I combined these factors into a so-called "perfect" double dose condition, I found that both the double dose effects became *positive*. This indicates that the choices in how to implement double dose can be as important as the choice to have double dose at all. Perhaps more importantly, very few "double dose" schools are implementing their policies in this way.

Although only a beginning, this study points the way to future research and policy. Under various forms, "double dose" is being adopted across the country with little research to support it. These policies are expensive in both time and money, and their outcomes are under-studied at best. This analysis suggests that although double dose may be beneficial under certain circumstances, it may even be *harmful* under less ideal implementation choices. More research is necessary to examine the different forms of implementation and their association with student achievement. Such research will be invaluable in deciding whether the input of resources is worth the benefits that may accrue.

CHAPTER V

DISCUSSION

The three sets of analyses presented here have each attempted to "Unpack Tracking" and provide entry points for policymakers to support low-achieving students. In Chapter II, I showed that gaps between high- and regular-track students persist in these large urban districts, and high-track students were substantially more likely to be exposed to high-status knowledge than their regular-track counterparts. The instructional quality measure used in this analysis had only a small relationship with student achievement on state standardized tests, however, so this difference mediated only a small portion of the relationship between track level and achievement. Although I recommend looking at other measures of instruction to see what is causing the gap in achievement gains between track levels, the rationing of high-status knowledge that is shown here is still a matter for great concern.

In Chapter III, I found that teachers' views of student ability are associated with higher achievement on standardized tests, and that holding a developmental view is particularly important in untracked settings. Teachers who see student ability as something that can be influenced by instruction and teachers who describe supports for struggling students that continue to engage them in rigorous mathematics on average have students with higher achievement. When students in untracked settings have teachers who hold productive views of supports for struggling students, they are

predicted to out-score tracked students, indicating that these views can support the success of untracked settings.

Finally, in Chapter IV, I examined the outcomes of one increasingly popular policy for supporting low-achieving students in middle school mathematics: double dose instruction. While the justification for embracing these policies seems to be that greater instructional time is associated with increased learning, in practice programs under the name of "double dose" varied significantly across schools and districts, and many of these variations had a significant impact on the relationship between double dose and achievement. Overall, double dose policies were associated with lower student achievement, both for the students in the double dose classes, and for the higher-achieving students in the same schools. However, the addition of school fixed effects showed that this difference could be attributed to unmeasured pre-existing differences between schools. As it turns out, school-level decisions in how to implement double dose policies were more important than the choice to adopt the policy itself. In fact, under the "perfect" implementation choices, double dose was associated with *higher* student achievement, although there were only four schools that chose to implement double dose in this way.

The findings across these three chapters indicate some avenues for policymakers and researchers alike in addressing modern tracking in middle school mathematics. First, to narrow the gaps in long-term outcomes between students, all students must be exposed to mathematics instruction that challenges them to reason and justify, rather than simply calculate and report. Second, for heterogeneous grouping of students to succeed, teachers must believe that all students can succeed in this type of mathematics instruction with the

correct supports. Finally, supports provided outside of the regular classroom, such as double dose instruction, have the potential to actually harm both low- and high-achieving students if implementation choices are not considered. Further research is needed in each of these areas, but these analyses point us toward a road to improving both short- and long-term outcomes for traditionally low-performing students.

REFERENCES

Abu el Haj, T.R. & Rubin, B.C. (2009). Realizing the Equity-Minded Aspirations of Detracking and Inclusion: Toward a Capacity-Oriented Framework for Teacher Education, *Curriculum Inquiry, 39*(3): 435-463.

Applebee, A.N., Langer, J.A., Nystrand, M. and Gamoran, A. (2003). Discussion-Based Approaches to Developing Understanding: Classroom Instruction and Student Performance in Middle and High School English. *American Educational Research Journal, 40*(3): 685-730.

Archbald, D., Glutting, J. and Qian, X. (2009). Getting into Honors or Not: An Analysis of the Relative Influence of Grades, Test Scores, and Race on Track Placement in a Comprehensive High School, *American Secondary Education, 37*(2): 65 – 81.

Baker, D. P., Fabrega, R., Galindo, C. and Mishook, J. (2004). Instructional Time and National Achievement: Cross-National Evidence, *Prospects, 34*(3): 311 – 334.

Balfanz, R., Letgers, N. and Jordan, W. (2004). Catching Up: Effect of the Talent Development Ninth-Grade Instructional Interventions in Reading and Mathematics in High-Poverty High Schools, National Association of Secondary School Principals, *NASSP Bulletin, 88*(641): 3 – 30.

Baron, R.M. and Kenny, D.A. (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic and Statistical Considerations, *Journal of Personality and Social Psychology, 51*(6): 1173 – 1182.

Bellei, C. (2009). Does Lengthening the School Day Increase Students' Academic Achievement? Results from a natural experiment in Chile*, Economics of Education Review, 28*(2009): 629 – 640.

Bloom, B.S. (1974). Time and Learning, *American Psychologist, 29*: 682 – 688.

Boaler, J. (2006). How a Detracked Mathematics Approach Promoted Respect, Responsibility, and High Achievement, *Theory into Practice, 45*(1): 40-46.

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.

Boston, M. (2012). Assessing Instructional Quality in Mathematics, *The Elementary School Journal, 113*(1): 76-104.

Boston, M. and Wolf, M.K. (2006). Assessing Academic Rigor in Mathematics Instruction: The Development of the Instructional Quality Assessment Toolkit, *CSE Technical Report 672,* Center for the Study of Evaluation.

Braddock, J.H. and Dawkins, M.P. (1993). Ability Grouping, Aspirations, and Attainments: Evidence from the National Educational Longitudinal Study of 1988, *The Journal of Negro Education, 62*(3): 324-336.

Brady, K. and Woolfson, L. (2008). What Teacher Factors Influence Their Attributions for Children's Difficulties in Learning? *British Journal of Educational Psychology, 78*(4): 527-544.

Brewer, D.J., Rees, D.I., Argys, L.M. (1995). Detracking America's Schools: The reform without cost? *Phi Delta Kappan, 77*(3): 210-214.

Brookings Institution (2009). How Well are American Students Learning? *The Brown Center Report on American Education.*

Brown, B.W. and Saks, D.H. (1986). Measuring the Effects of Instructional Time on Student Learning; Evidence from the Beginning Teacher Evaluation Study, *American Journal of Education, 94*(4): 480 – 500.

Brown, M.A. and Gray, M.W. (1992). Mathematics Test, Numerical, and Abstraction Anxieties and Their Relation to Elementary Teachers' Views on Preparing Students for the Study of Algebra, *School Science and Mathematics, 92*(2): 69 – 73.

Burris, C.C., Wiley, E., Welner, K., 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.

Carbonaro, W. (2005). Tracking, students' effort, and academic achievement. *Sociology of Education, 78*, 27–49.

Carbonaro, W.J. and Gamoran, A. (2002). The production of Achievement Inequality in High School English, *American Educational Research Journal, 39*(4): 801-827

Cavanagh, S. (2006). Students Double-Dosing on Reading and Math: Schools aim to improve state test scores—and satisfy federal education law, *Education Week, 25*(40): 1 – 2.

Chait, R., Muller, R. D., Goldware, S., & Housman, N. G. (2007). *Academic interventions to help students meet rigorous standards: State policy options*.

Washington, D.C.: The National High School Alliance at the Institute for Educational Leadership.

Cobb, J., Boufi, A., McClain, K. and Whitenack, J. (1997). Reflective Discourse and Collective Reflection*, Journal for Research in Mathematics Education, 28*(3): 258-277.

Cohen, E.G. (1972). *Designing Groupwork: Strategies for the heterogeneous classroom*, Teachers College Press: New York, NY.

Cohen, E.G. and Lotan, R.A. (1997). *Working for Equity in Heterogeneous Classrooms*, Teachers College Press: New York, NY.

Council of the Great City Schools (2009). *Urban Indicator: High School Reform Survey.* Retrieved October, 2011 from http://cgcs.schoolwires.net/cms/lib/DC00001581/Centricity/Domain/35/Publication%20Docs/Urban_Indicator09.pdf

Darling-Hammond, L. (1998). Teachers and Teaching: Testing Policy Hypothesis from a National Commission Report, *Educational Researcher, 27*(1): 5- 15.

Dreeben, R. and Barr, R. (1988). Classroom Composition and the Design of Instruction, *Sociology of Education, 61*(3): 129-142.

Dupriez, V., Dumay, X. and Vause, A. (2008). How Do School Systems Manage Pupils' Heterogeneity? *Comparative Education Review, 52*(2): 245 – 273.

Eder, D. (1981). Ability Grouping as a Self-Fulfilling Prophesy: A micro-analysis of teacher-student interaction, *Sociology of Education, 54*(July): 151-162.

Epstein, J.L. & MacIver, D.J. (1992*). Opportunities to learn: Effects on eighth graders of curriculum offerings and instructional approaches (Report No. 34).* Baltimore, MD: Center for Research on Effective Schooling for Disadvantaged Students.

Esposito, D. (1973). Homogeneous and Heterogeneous Ability Grouping: Principal Findings and Implications for Evaluating and Designing More Effective Educational Environments, *Review of Educational Research, 43*(2): 163 – 179.

Evertson, C.M. (1982). Differences in Instructional Activities in Higher- and Lower-Achieving Junior High English and Math Classes, *The Elementary School Journal, 82*(4): 329 – 350.

Finley, M.K. (1984). Teachers and Tracking in a Comprehensive High School, *Sociology of Education, 57*(4): 233-243.

Freeman, B. and Crawford, L. (2008). Creating a Middle School Mathematics Curriculum for English-Language Learners, *Remedial and Special Education, 29*(1): 9 – 19.

Gamoran, A. (1986). Instructional and Institutional Effects of Ability Grouping, *Sociology of Education, 59*(4): 185-198.

Gamoran, A. (1987). The Stratification of High School Learning Opportunities, *Sociology of Education*, 60(3): 135-155

Gamoran, A. (2009). Tracking and Inequality: New Directions for Research and Practice. *WCER Working Paper Number 2009-6* Retrieved October 2012 from http://www.wcer.wisc.edu/

Gamoran, A. and Weinstein, M. (1998). Differentiation and Opportunity in Restructured Schools. *American Journal of Education 106*: 385-415.

Gamoran, A., Nystrand, M, Berends, M. and LePore, P.C. (1995). An Organizational Analysis of the Effects of Ability Grouping, *American Educational Research Journal, 32*(4): 687-715.

Giglio, K. (2010). What Teacher Characteristics Affect Student Achievement? Findings from Los Angeles Public Schools, *RAND Research Brief,* Retrieved April 2012 from http://www.rand.org.proxy.library.vanderbilt.edu/pubs/research_briefs/2010/RAND_RB9526.pdf

Gould, S. (1981). *The Mismeasure of Man*, New York, NY: Norton Publishing.

Hallinan, M. (1994). Tracking: From Theory to Practice, *Sociology of Education, 67*(2): 79-91.

Hallinan, M.T. (1994). School Differences in Tracking Effects on Achievement, *Social Forces, 72*(3): 799-820.

Hand, V.M. (2010). Co-Construction of Opposition in a Low-Track Mathematics Classroom, *American Educational Research Journal, 47*(1): 97 – 132.

Hanushek, E. A., & Woßmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *Economic Journal, 116*(510), C63–C76.

Henningsen, M. and Stein, M.K. (1997). Mathematical Tasks and Student Cognition: Classroom-Based Factors that Support and Inhibit High-Level Mathematical Thinking and Reasoning, *Journal for Research in Mathematics Education, 28*(5): 524-549.

Heubert, J. P., & Hauser, R. M. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Research Council.

Horn, I.S. (2006). Lessons Learned from Detracked Mathematics Departments. *Theory into Practice, 45*(1): 72-81.

Horn, I.S. (2007). Fast Kids, Slow Kids, Lazy Kids: Framing the Mismatch Problem in Mathematics Teachers' Conversations, *Journal of the Learning Sciences*, 16(1): 37 – 79.

Jackson, K., Gibbons, L., Garrison, A. and Munter, C. (2012). *Exploring Relationships Mathematics Teachers' Views of Students' Capabilities, Visions of Instruction, and Instructional Practices*, Unpublished Manuscript.

Junker, B. and Weisberg, Y. Matsumura, L.D., Crosson, A., Wolf, M.K., Levison, A. and Resnick, L. (2006). Overview of the Instructional Quality Assessment, *CSE Technical Report 671,* Center for the Study of Evaluation.

Kemple, J. J., Herlihy, C. M., & Smith, T. J. (2005). *Making progress toward graduation: Evidence from the talent development high school model*. New York: MDRC.

Kenny, D.A. (2008). *Mediation with Dichotomous Outcomes,* Retrieved December 2012 from http://davidakenny.net/doc/dichmed.pdf

Krull, J.L. and MacKinnon, D.P. (1999). Multilevel Mediation Modeling in Group-Based Intervention Studies, *Evaluation Research, 23*(4): 418 – 444.

Krull, J.L. and MacKinnon, D.P. (2001). Multilevel Modeling of Individual and Group Level Mediated Effects, *Multivariate Behavioral Research, 36*(2): 249 – 277

Kukla-Acevedo, S. (2009). Do Teacher Characteristics Matter? New Results on the Effects of Teacher Preparation on Student Achievement, *Economics of Education Review, 28*(1): 49-57.

Kulik, C.C. and Kulik, J.A. (1982). Effects of Ability Grouping on Secondary School Students: A Meta-Analysis of Evaluation Findings, *American Educational Research Journal, 19*(3): 415-428

Lotan, R. (2006). Teaching Teachers to Build Equitable Classrooms, *Theory into Practice, 45*(1): 32 – 39.

Loveless, T. (1999). *The tracking wars: State reform meets school policy*. Washington, DC: Brookings Institution Press.

Loveless, T. (2009). *Tracking and Detracking: High achievers in Massachusetts middle schools*. Dayton, OH: Fordham Institute Press.

Lucas, S.R. (1999). *Tracking Inequality: Stratification and Mobility in American High Schools*. New York, NY: Teachers College Press

Lucas, S.R. and Berends, M. (2002). Sociodemographic Diversity, Correlated Achievement, and De Facto Tracking, *Sociology of Education, 75*(4): 328-348.

Mac Iver, D.J. (1991). Helping Students Who Fall Behind: Remedial Activities in the Middle Grades, *Department of Education Report Number 22*, October 1991.

MacKinnon, D. P., & Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review, 17*, 144-158.

Marcotte, D.E. (2007). Schooling and Test Scores: A mother-natural experiment, Economics of *Education Review, 26*(2007): 6219 – 640.

Matsumura, L.C., Garnier, H.E., Slater, S.C., & Boston, M.D. (2008). Toward Measuring Instructional Interactions "At Scale," *Educational Assessment, 13*(4): 267 – 300.

McDermott, P., Rothenberg, J. & Martin, G. (1995). "Should We Do It the Same Way?" Teaching in Tracked and Untracked High School Classes. *Paper Presented at the Northeastern Educational Research Association 26th Annual Conference*, October 25-27, 1995

Mosteller, F., Light, R.J., Sachs, J.A. (1996). Sustained Inquiry in Education: Lessons from Skill Grouping and Class Size, *Harvard Educational Review, 66*(4): 797-842

National Council of Teachers of Mathematics. (2000*). Principles and Standards for School Mathematics*. Reston, VA: Author.

Ney, R.J. (2010). *A Study of Doubling Class Time for Low Achieving High School English and Math Students and the Impact on State Tests Required Under NCLB*, Unpublished Doctoral Dissertation, Liberty University, Lynchburg, VA.

Nomi, T. and Allensworth, E. (2008). *"Double Dose" Algebra as an Alternative Strategy to Remediation: Effects on Students' Academic Outcomes*, Working Paper, Chicago Consortium on School Research.

Nomi, T. and Allensworth, E. (2011). Sorting and Supporting: Why Double-Dose Algebra Led to Better Test Scores but More Course Failure through Changes in Classroom Composition, Climate and Instruction*, Paper Presented at the 2011 Annual Meeting of the American Educational Research Association*, New Orleans, LA.

Oakes, J. (1982). The Reproduction of Inequity: The content of secondary school tracking, *The Urban Review, 14*(2): 107- 120.

Oakes, J. (2005*). Keeping Track: How schools structure inequality*, Second Edition. New Haven, CT: Yale University Press.

Oakes, J., Wells, A.S., Jones, M. & Datnow, A. (1997). Detracking: The Social Construction of Ability, Cultural Politics, and Resistance to Reform, *Teachers College Record, 98*(Spring): 482 – 510.

Page, R. N. (1991). *Lower track classrooms: A curricular and cultural perspective*. New York: Teachers College Press.

Pearson (2012). *SuccessMaker Software*, Retrieved December 2012 from http://www.pearsonschool.com/index.cfm?locator=PSZk99

Peele, L.L. (1998). Double-Dose: A Viable Instructional Alternative*, National Association of Secondary School Principals Bulletin*, 82(599): 111 – 114.

Perkins-Gough, D. (2006). Accelerating the Learning of Low-Achievers, *Educational Leadership*, 63(5): 88 – 89.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods second edition*. London: Sage.

Reed, J. (2008). Shifting Up: A Look at Advanced Mathematics Classes in Tracked Schools. *The High School Journal,* 91(4): 45 – 58.

Resnick, D.P. & Resnick, L.B. (1985). Standards, Curriculum, and Performance: A Historical and Comparative Perspective. *Educational Researcher, 14*(4) 5-20.

Resnick, L., Matsumura, C. and Junker, B. (2006*).* Measuring Reading Comprehension and Mathematics Instruction in Urban Middle Schools: A Pilot Study of the Instructional Quality Assessment, *CSE Technical Report 681,* Center for the Study of Evaluation.

Rist, R.C. (1970). Student Social Class and Teacher Expectations: The Self-Fulfilling Prophesy in Ghetto Education, *Harvard Educational Review, 40*(3): 411- 451.

Rosenbaum, J.E. (1976). *Making inequality: The hidden curriculum of high school tracking*. New York, NY: John Wiley & Sons.

Rubin, B. (2008). Detracking in Context: How Local Constructions of Ability Complicate Equity-Geared Reform. *Teachers College Record, 110*(3): 646-699.

Rubin, B.C. (2003). Unpacking Detracking: When Progressive Pedagogy Meets Students' Social Worlds. *American Educational Research Journal, 40*(2): 539-573.

Rubin, B.C. (2006). Tracking and Detracking: Debates, Evidence, and Best Practices for a Heterogeneous World, *Theory into Practice, 45*(1): 4 - 14.

Schmidt, R.A. (2011, April). The Relationship between Tracking and Mathematics Achievement Test Outcomes in Four Urban Districts. *Paper presented at the annual conference of the American Educational Research Association (AERA)*, New Orleans, LA.

Silva, E. (2007*). On the Clock: Rethinking the Way Schools Use Time*, Washington, DC: Education Sector.

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis, *Review of Educational Research, 60*(3): 471-499

Spielhagen, F.R. (2006). Closing the Achievement Gap in Math: Considering Eighth Grade Algebra for All Students, *American Secondary Education, 34*(3): 29-42.

StataCorp. (2009*). Stata Multiple Imputation Reference Manual (Release 11).* College Station, TX.

Stein, M.K., Grover, B.W. and Henningsen, M. (1996). Building Student Capacity for Mathematical Thinking and Reasoning: An Analysis of Mathematical Tasks Used in Reform Classrooms, *American Educational Research Journal, 33*(2): 455 - 488.

Villegas, A.M. (1991). Life in the Classroom: The Influence of Track Placement and Student Race/Ethnicity. In *On the Right Track: The Consequences of Mathematics Course Placement Policies and Practices in the Middle Grades.* Chicago, IL: American Educational Research Association.

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.

Watanabe, M., Nunes, N., Mebane, S., Scalise, K. & Claesgens, J. (2007). "Chemistry for All, Instead of Chemistry Just for the Elite": Lessons Learned from Detracked Chemistry Classrooms, *Wiley InterScience*, www.interscience.wiley.com

Wheelock, Anne. (1992). *Crossing the Tracks: How "Untracking" Can Save America's Schools*. New York, NY: New Press.

Worthy, J. (2010). Only the Names Have Been Changed: Ability Grouping Revisited, *Urban Review, 42*: 271-295.

Zhang, Z., Zyphur, M.J., and Preacher, K.J. (2008). Testing Multilevel Mediation Using Hierarchical Linear Models, *Organizational Research Methods, 12*(4): 695-719.

Zohar, A., Degani, A. and Vaaknin, E. (2000). Teachers' Beliefs about Low-Achieving Students and Higher Order Thinking*, Teaching and Teacher Education, 17*(2001): 469-485.