Essays in the Economics of Education

by

Jonathan T Schellenberg

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Christopher Walters, Chair
Professor David Card
Professor Jesse Rothstein

Spring 2019

Essays in the Economics of Education

Abstract

Essays in the Economics of Education

by

Jonathan T Schellenberg

Doctor of Philosophy in Economics

University of California, Berkeley

Associate Professor Christopher Walters, Chair

Investments in human capital can have large and long-lasting impacts on students. This dissertation studies the relationship between early education and long-run outcomes of students, with a particular focus on future criminal behavior, and examines how teacher quality and school choice influence these future gains.

My first chapter, which is joint work with Evan K. Rose and Yotam Shem-Tov, investigates the impact of teacher quality on future criminal behavior. Using a unique data set linking the universe of public school records to administrative criminal justice records for the state of North Carolina, we demonstrate strong associations between future criminal activity and early life education outcomes including test scores, attendance, and disciplinary records. We estimate value-added models measuring the causal impacts of teachers on short-run cognitive and non-cognitive outcomes in a multivariate random effects framework, and link these short-run effects to teacher effects on adult crime. We find that teachers primarily influence future crime through a non-cognitive channel, and that their cognitive and non-cognitive impacts are orthogonal. This result implies that test score-based measures miss an important component of the social value of teacher quality, suggesting scope for improved teacher assessment systems that also account for non-cognitive gains.

I build on the relationship between early life education and crime in my second chapter, which studies the explanatory power of educational achievement on the black-white gap in criminal offending rates. We document strong relationships between test scores and future criminality. We show that observable differences between blacks and whites in early grades, including neighborhoods, schools, and other demographic information, can explain the differences in their relative rates of being charged with any offense in early adulthood, and that test score differences can explain between a quarter to a half of this gap. This difference in offending is akin to the "skill gap" described by Neal and Johnson (1996) that explains a large fraction of the raw black-white gap in wages. We also document two important nuances to this story. First, while observable differences can explain nearly the entire gap in charge rates for *any* offense, we still are unable to explain about a quarter of the difference in felony offending rates. Second, we show that blacks experience a much greater return to skill than

white students in the form of reduced crime, and that these differential returns explain a substantial fraction – between 10% and 20% – of the raw crime gap. This difference in returns to higher achievement is particularly relevant for more severe offenses, and plays a larger role in explaining the differences in offending rates between black and white men from worse economic backgrounds.

My third chapter is based on joint work with Atila Abdulkadiroğlu, Parag Pathak, and Christopher Walters, and studies how school choice affects parents' educational investments for their children. We study relationships among parent preferences, peer quality, and causal effects on outcomes for applicants to New York City's centralized high school assignment mechanism. We use applicants' rank-ordered choice lists to measure preferences and to construct selection-corrected estimates of treatment effects on test scores, high school graduation, college attendance, and college quality. Parents prefer schools that enroll high-achieving peers, and these schools generate larger improvements in short- and long-run student outcomes. Preferences are unrelated to school effectiveness and academic match quality after controlling for peer quality.

To Ethan, Kyle, and Tristan, who helped made this possible.

Contents

Contents **ii**

List of Figures **iv**

List of Tables **v**

1 The Effects of Teacher Quality on Crime **1**

1.1 Introduction 1
1.2 Data and Settings 4
1.3 Relationship Between Education and Criminality 6
1.4 Estimation of Teacher Effects 9
1.5 Results from Univariate Model 14
1.6 Multivariate Random Coefficients Approach 18
1.7 Policy Consequences 21
1.8 Conclusion 23
1.9 Figures 23
1.10 Tables 28
1.11 Appendix: Data Construction 34
1.12 Appendix: Relationship Between 2SLS and Value-Added 36
1.13 Appendix: Additional Figures 39
1.14 Appendix: Additional Tables 46

2 Can Racial Gaps in Offending Be Explained? **58**

2.1 Introduction 58
2.2 Data 60
2.3 Observed Relationship Between Test Scores and Crime 61
2.4 Decomposition of the Race Gap Using Test Scores 66
2.5 Discussion and Conclusion 68
2.6 Figures 69
2.7 Tables 73
2.8 Appendix: Additional Figures 78
2.9 Appendix: Additional Tables 83

3 Do Parents Value School Effectiveness? 96

3.1 Introduction 96
3.2 Setting and Data 99
3.3 Conceptual Framework 102
3.4 Empirical Methods 103
3.5 Parameter Estimates 109
3.6 Preferences, Peer Quality, and School Effectiveness 112
3.7 Discussion 115
3.8 Conclusion 117
3.9 Figures 118
3.10 Tables 121
3.11 Appendix: Data 132
3.12 Appendix: Econometric Methods 134
3.13 Appendix: Proof of Identity to Express Control Function Non-Recursively . 139
3.14 Appendix: Additional Tables 143

Bibliography 154

List of Figures

1.1 Relationships Between Grade 3 Test Scores and Criminal Charge Rates 24
1.2 Relationships Between Test Scores and Age 20 Crime Rates 25
1.3 Relationships Between Test Scores and Age 20 Crime Rates By Demographics . 26
1.4 Relationships Between Behavioral Measures and Age 20 Crime Rates 27
1.5 Relationships Between Test Scores and Crime Rates By Age and Crime Severity 40
1.6 Relationships Between Test Scores and Crime Rates by Age and Crime Type . . 41
1.7 Relationships Between Test Scores and Age 20 Felony Charge Rates Across Demographic Groups 42
1.8 Relationships Between Test Scores and Age 20 Assault Charge Rates Across Demographic Groups 43
1.9 Relationships Between Test Scores and Age 20 Property Crime Charge Rates Across Demographic Groups 44
1.10 Relationships Between Test Scores and Age 20 Drug Crime Rates Across Demographic Groups 45
1.11 Binned Scatter Plots of Value-Added Measures 46

2.1 Relationships Between Test Scores and Criminal Charge Rates by Race 70
2.2 Relationships Between Test Scores and Crime by Crime Severity and Race . . . 71
2.3 Relationships Between Test Scores and Criminal Charge Rates by Race and Socio-Economic Status 72
2.4 Relationships Between Test Scores and Criminal Charge Rates by Race and Type of Criminal Offense 78
2.5 Relationships Between Test Scores and Criminal Charge Rates by Race and Gender 80
2.6 Relationships Between Test Scores and Criminal Charge Rates by Race and Gender for Economically Disadvantaged Students 81
2.7 Relationships Between Test Scores and Criminal Charge Rates by Race and Gender for Economically Non-Disadvantaged Students 82

3.1 Comparison of Value-Added and Control Function Estimates of School Average Treatment Effects 119
3.2 Relationships between effects on test scores and effects on long run outcomes . . 120
3.3 Relationships among preferences, peer quality, and Regents math effects 121

List of Tables

1.1 Summary Statistics 28
1.2 Distribution of Teacher Value-Added 29
1.3 Impacts of Value-Added on Crime 30
1.4 Multivariate Distributions of Value-Added 31
1.5 Impacts of Total Elementary Value-Added on Criminal Activity at Age 20 . . . 32
1.6 Policy Simulations 33
1.7 Summary Statistics for Middle School Sample 47
1.8 Estimating Forecast Bias in Value-Added Estimates 48
1.9 Impacts of Value-Added on Criminal Activity by Race 49
1.10 Impacts of Value-Added on Criminal Activity by Family Economic Status . . . 50
1.11 Pooled Impacts of Value-Added on Criminal Activity by Age of Outcome 51
1.12 Impacts of Value-Added on Criminal Activity by Teacher Grade 52
1.13 Value-Added Measures Using Combinations of Elementary School Teacher Effects 53
1.14 Impacts of Value-Added on Crime Outcomes Using Alternative Shrinkage Procedures 54
1.15 Impacts of Value-Added on Criminal Activity for Middle School Teachers 55
1.16 Impacts of Value-Added on Criminal Activity at Age 18 for North Carolina High School Students 56
1.17 Impacts of Value-Added on Criminal Activity at Age 20 for North Carolina High School Students 57

2.1 Summary Statistics 73
2.2 The Residual Black-White Crime Gap in All Charges 74
2.3 Residual Black-White Crime Gaps 75
2.4 Residual Black-White Crime Gaps for Economically Disadvantaged Students . . 75
2.5 Residual Black-White Crime Gaps for Economically Non-Disadvantaged Students 76
2.6 Blinder-Oaxaca Decomposition of the Black-White Crime Gap at Age 20 77
2.7 Blinder-Oaxaca Decomposition of the Black-White Crime Gap by Economic Status 79
2.8 Summary Statistics for Students With Observed Convictions 83
2.9 The Residual Black-White Crime Gap in Charges by Age 20 Using Different Controls and Only Grade 8 Test Scores 84

2.10 Magnitude of the Residual Black-White Gap in the NLSY97 for Arrests by Age 20 Using Different Controls . . . 85
2.11 Magnitude of the Residual Black-White Gap in the NLSY97 for Incarceration by Age 20 Using Different Controls . . . 86
2.12 The Residual Black-White Crime Gap in Charges by Age 20 When Controlling for Educational Attainment . . . 87
2.13 Magnitude of the Residual Black-White Crime Gap at Age 20 By Crime Type and Types of Test Score Controls . . . 88
2.14 Magnitude of the Residual Black-White Crime Gap for Men at Age 20 By Charge Severity and Types of Test Score Controls . . . 88
2.15 Magnitude of the Residual Black-White Crime Gap for Women at Age 20 By Charge Severity and Types of Test Score Controls . . . 89
2.16 Blinder-Oaxaca Decomposition of the Black-White Crime Gap in Misdemeanors at Age 20 Using Test Scores in Different Years . . . 90
2.17 Blinder-Oaxaca Decomposition of the Black-White Crime Gap in Felonies at Age 20 Using Test Scores in Different Years . . . 91
2.18 Blinder-Oaxaca Decomposition of the Black-White Crime Gap in Convictions at Age 20 Using Test Scores in Different Years . . . 92
2.19 Blinder-Oaxaca Decomposition of Test Scores and the Black-White Crime Gap for Males at Age 20 . . . 93
2.20 Blinder-Oaxaca Decomposition of Test Scores and the Black-White Crime Gap at Age 20 By Crime Type and Economic Status for Male Students . . . 94
2.21 Blinder-Oaxaca Decomposition of Eighth Grade Math Test Scores Only and the Black-White Crime Gap at Age 20 . . . 95

3.1 Descriptive Statistics for New York City Eighth Graders . . . 122
3.2 Correlates of Preference Rankings for New York City High Schools . . . 123
3.3 Variation in Student Preference Parameters . . . 124
3.4 Distributions of Peer Quality and Treatment Effect Parameters for Regents Math Scores . . . 125
3.5 Correlations of Peer Quality and Treatment Effect Parameters for Regents Math Scores . . . 126
3.6 Decomposition of School Average Outcomes . . . 127
3.7 Preferences for Peer Quality and Regents Math Effects . . . 128
3.8 Preferences for Peer Quality and School Effectiveness by Outcome . . . 129
3.9 Heterogeneity in Preferences for Peer Quality and Regents Math Effects . . . 130
3.10 Potential Achievement Gains from Ranking Schools by Effectiveness . . . 131
3.11 Sample Restrictions . . . 144
3.12 Correlations of Peer Quality and Treatment Effect Parameters for Regents Math Scores, Value-Added Model . . . 145
3.13 Joint Distribution of Peer Quality and Treatment Effect Parameters for PSAT Scores/10 . . . 146

3.14 Joint Distribution of Peer Quality and Treatment Effect Parameters for High School Graduation 147
3.15 Joint Distribution of Peer Quality and Treatment Effect Parameters for College Attendance 148
3.16 Joint Distribution of Peer Quality and Treatment Effect Parameters for Log College Quality 149
3.17 Preferences, Peer Quality, and Math Effects, Alternative Measures of Popularity 150
3.18 Preferences for Peer Quality and Regents Math Effects Among Students Ranking Fewer Than 12 Choices 151
3.19 Preferences, Peer Quality, and Math Effects, Alternative Treatment Effect Models 152
3.20 Potential Achievement Gains from Ranking Schools by Effectiveness, by Baseline Test Score Quartile 153

Acknowledgments

First and foremost, I am extremely grateful to my incredible advisor, Chris Walters, for his continuous support over my graduate school career. Chris went out of his way to provide opportunities for me to succeed and worked tirelessly as my guide and mentor. I greatly appreciate his generous efforts to assist me and his superhuman patience.

I am also extremely thankful to the members of my dissertation committee, David Card and Jesse Rothstein, for their extensive academic support and guidance. Their efforts have had a profound impact on my work. I would also like to thank the other faculty in the department for their assistance, in particular Hilary Hoynes, Pat Kline, Conrad Miller, and Enrico Moretti, who all provided key insights and very helpful comments at every stage of these projects.

I have been fortunate enough to collaborate with an amazing group of co-authors, namely Evan K. Rose, Yotam Shem-Tov, Atila Abdulkadiroğlu and Parag Pathak. None of this work would be possible without them. Evan and Yotam worked tirelessly to secure the data used in this dissertation and have provided invaluable insights into all of my work presented here. Working with Atila and Parag has been an amazing experience, and I am extremely grateful that I had the opportunity to do so – I learned so much from them.

My progression through graduate school has been greatly enhanced by the support and assistance from my classmates and friends. My research has been directly improved by feedback from Natalie Bachas, Zarek Brot-Goldberg, Christina Brown, Alessandra Fenizia, Jonathan Holmes, Sheisha Kulkarni, Jennifer Kwok, Julien Lafortune, Nicholas Li, Juliana Londoño-Vélez, Ian Luby, Waldo Ojeda, Kai Peterson, Deepak Premkumar, Raffaele Saggio, Avner Strulov-Shlain, and Kevin Todd, and my time in graduate school has been enriched by their presence.

I also would like to acknowledge the people in my life outside of academia that have helped me throughout this process. My parents and my brother Geoff have been extremely supportive through the ups and downs of graduate school. I also greatly appreciate the support I received from Abe, Ashley, Daeus, Erik, Ganesh, Kyle, Sam, and many more.

I would like to acknowledge the North Carolina Education Research Data Center and the New York Department of Education for providing the data used in my research. The National Academy of Education / Spencer Foundation, along with Alan Auerbach, David Card, Jesse Rothstein, and Emmanuel Saez, provided the funding necessary for these projects, and I am grateful for their support.

Finally, a big thank you to UC Berkeley for creating a community that I could call home for the past six years. It has been a wonderful journey.

Chapter 1

The Effects of Teacher Quality on Crime

1.1 Introduction

The impact of teacher quality on future adult outcomes is an important policy question that remains controversial among scholars and policymakers. Teacher quality is most commonly evaluated using test score "value-added" (VA), which captures test score gains conditional on student-level observables. Recent work has documented large variation in teacher quality, and that teachers with high test score VA also improve their students' post-secondary outcomes, reduce teenage birth rates, and increase earnings (Chetty, Friedman, and Rockoff, 2014b). These findings have added motivation to the debate about using performance-pay incentives for teachers based on test score gains (Hanushek, 2011; Neal, 2011).

An additional adult outcome that teachers may influence but has received comparatively less attention in the VA literature is crime. Crime is an important outcome for evaluating the impacts of education, since it is widespread, concentrated in impoverished school districts, and often begins at an early age. In the United States, over a quarter of the population has been arrested by age 21 (Brame, Turner, Paternoster, and Bushway, 2012). Moreover, crime reduction is one of the largest sources of social returns from early-life educational investments (Heckman, Moon, Pinto, Savelyev, and Yavitz, 2010a). However, it is unclear if test score gains, the predominant methodology for measuring teacher quality, will capture teachers' impacts on criminal behavior. Given that non-cognitive skills are better predictors of behavioral outcomes, including crime (Heckman and Kautz, 2012; Heckman, Stixrud, and Urzua, 2006a), there is reason to believe that teachers' impacts on crime operate through a non-cognitive channel (Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan, 2011). Performance-pay incentives for teachers that rely on test score VA may therefore miss an important component of teacher effectiveness, especially if teachers' impacts on test scores are unrelated to their impacts on non-cognitive skills (Neal, 2011).

This paper estimates the impact of teacher quality on students' criminal outcomes as adults. Our work focuses directly on the mechanisms through which teacher quality affects future criminality. We estimate teacher VA models for short-run cognitive and non-cognitive outcomes in a multivariate random effects framework, and link these short-run effects to teacher effects on adult crime. We then conduct policy counterfactuals that compare teacher retention policies based on various performance measures.

To conduct our analyses, we linked administrative public school records in North Carolina to the universe of court records in the state. These data include rich student demographics, tract level address information, test scores, disciplinary and attendance records, and all criminal charges and convictions. The richness of our data allows us to study statistical relationships between early education and crime that have yet to be established using administrative records.[1]

We begin our analysis by establishing a new set of descriptive facts relating elementary and middle school educational achievement to future crime. Raw correlations show a very robust association between test scores and crime – a one standard deviation increase in third grade math (reading) scores is associated with a 2.25 (2.83) percentage point decrease in the likelihood of being charged with a crime by age 20, roughly 10% of the average charge rate. We also see strong correlations between early student behavior within the classroom and their future criminal behavior. For instance, a suspension in third grade is associated with a 19.0 percentage point increase in criminal charge likelihood by age 20, and a 10% increase in absences in third grade is associated with a 18.6 percentage point increase in the age 20 criminal charge rate. These associations capture a combination of selection bias and any causal effects that teachers may have on criminality.

To identify a causal relationship between improved teacher quality and future interactions with the criminal justice system, we evaluate teacher effectiveness for elementary school teachers using a VA framework, similar to prior work (Rockoff, 2004; Kane and Staiger, 2008; Chetty, Friedman, and Rockoff, 2014a; Rothstein, 2017). However, as noted above, test scores may not be sufficient to capture teacher effects on crime, given that childhood behavior is a strong predictor of future criminality (Carneiro, Crawford, and Goodman, 2007; Reynolds, Temple, and Ou, 2010)). To construct a measure of "non-cognitive VA", we follow Jackson (2018) and Petek and Pope (2016) and estimate teacher impacts on changes in behavioral outcomes. Our main measures of non-cognitive VA examine elementary school teachers' impacts on their students' future behavior, namely their truancy and propensity to receive disciplinary action in middle school (Petek and Pope, 2016). We avoid focusing on contemporaneous measures of behavior, particularly with suspensions, since teachers can directly impact their students' contemporaneous disciplinary records.

To account for sampling error in the estimates of teacher quality, researchers commonly use empirical Bayes shrinkage techniques. As typically implemented, this procedure imposes the assumption that student observables are independent of teacher effectiveness, a condition that we show to be violated in practice. We therefore introduce a conditional shrinkage

[1] Other studies have connected early cognitive achievement either through small-scale experiments (Heckman et al., 2010a) or through longitudinal survey data (Heckman et al., 2006a).

procedure that reduces the variance of individual teacher effects, while also accounting for any relationships between student characteristics and teacher effects.[2]

Our estimates of teachers' impacts show that teachers generate large impacts on both test scores and future behavioral outcomes in the short run. We then investigate how these short-run teacher effects are related to their students' future criminality. We find suggestive evidence that teachers who improve reading test scores also reduce crime. These effects are modest – using our preferred specification, a one standard deviation improvement in reading test score VA leads to a 0.1 percentage point decrease in their students' criminal charge rates at age 20, or about 0.4% relative to the mean charge rate. The effects of math VA on crime are smaller and statistically insignificant. However, our results are highly sensitive to specification, and we cannot assertively conclude that teachers' cognitive impacts lead to reduced crime.

In contrast, we find much more robust evidence that teachers' impacts on short-term behavioral outcomes have long-run effects on their propensity to commit crime. Students with teachers that lead to a standard deviation higher probability of being suspended in grade 6 are also more likely to commit a crime by 0.38 percentage points. Similarly, teachers who increase their students' future truancy increase their criminality. These findings are robust to a number of specifications. This evidence shows that teachers' direct impacts on behavioral outcomes have large, long-run consequences for the student.

Our results establish that high quality teachers reduce crime, and that these effects on crime may operate through both cognitive and non-cognitive channels. Motivated by these findings, we build a multivariate random effects framework to estimate the distribution of teacher effects along multiple dimensions of teacher quality.[3] We then use this two-dimensional model of teacher effectiveness to test for differential impacts of teacher quality on their students' future crime. Our estimates show that teacher impacts on test scores and future disciplinary records are unrelated to one another, and that the teacher's latter component of ability is the only dimension through which teachers can impact criminal behavior.

Finally, we consider the implications of our findings for the design of teacher personnel policies that aim to reduce crime. We find that replacing teachers in the bottom 5% of crime deterrence with median teachers leads to a 0.27% decrease in all charges by age 20, with larger effects (0.55%) for felonies. Moreover, we find that these returns are entirely determined through a non-cognitive channel, and that hiring policies with a non-cognitive component captures over 50% of the maximal feasible reduction. Thus, test score VA of teacher evaluations is not sufficient to minimize crime. The non-cognitive channel of teacher efforts have meaningful effects that are orthogonal to cognitive outcomes, indicating that

[2] To our knowledge, conditional shrinkages have not been used in education papers. These methods are used in recent estimates of health care quality (Chandra, Finkelstein, Sacarny, and Syverson, 2016).

[3] This approach is relatively uncommon in the VA literature. Broatch and Lohr (2012) consider a multivariate approach, although in our practice, we will consider the joint distribution of a continuous and binary outcome as opposed to all continuous and all binary. Chamberlain (2013) use a similar multivariate approach to separate teacher effects on test scores and college outcomes.

teacher evaluation solely on test scores is suboptimal.

Our study adds to the burgeoning literature highlighting the relationship between education and crime. Recent studies have found that many different aspects of the educational system, including compulsory schooling (Lochner and Moretti, 2004; Cook and Kang, 2016; Jacob and Lefgren, 2003), redistricting (Billings, Deming, and Rockoff, 2013), and enrollment at more desirable schools (Deming, 2011), affect students' likelihood of future criminality. This paper highlights another scope for crime reduction through the education system: improved teacher effectiveness.

A common theory used to explain crime reduction through improved education is a non-cognitive skill development mechanism (Cunha, Heckman, and Schennach, 2010; Lochner, 2011). In line with this finding, longitudinal studies of early educational interventions have shown long-lasting reductions in crime of their participants, and that these returns are unrelated to short-lived test score gains (Anderson, 2008; Heckman et al., 2010a; Heckman, Moon, Pinto, Savelyev, and Yavitz, 2010b). Our work provides direct evidence of crime reduction through non-cognitive channel by linking teachers' short-run effects on student behavior to their long-run criminal outcomes.

A separate and extensive literature has investigated the long-run effects of school (Abdulkadiroğlu, Angrist, Dynarski, Kane, and Pathak, 2011; Abdulkadiroğlu, Pathak, Schellenberg, and Walters, 2017b) and teacher (Jackson, Rockoff, and Staiger, 2014; Chetty et al., 2014b) quality, as measured using value-added scores. Recent work studying teacher quality highlights teachers' non-cognitive impacts on behavioral outcomes, including high school completion and college intentions (Jackson, 2018; Petek and Pope, 2016; Flèche, 2017; Gershenson, 2016). We extend the VA literature to connect teacher effects, both cognitive and non-cognitive, to future criminality, and to rigorously estimates teachers' impacts across multiple dimensions.

The order of the rest of this paper is as follows. Section 1.2 describes our education and crime records for North Carolina. Section 1.3 will provide a descriptive analysis showing the relationships between early education and crime. Section 1.4 will outline our empirical approach to estimating teacher effects, and Section 1.5 will summarize these results. Section 1.6 will modify the prior approach to estimate multidimensional teacher effects. Section 1.7 estimates crime reduction under alternative teacher hiring policies. Section 1.8 concludes.

1.2 Data and Settings

Education Records

We utilize administrative education records, provided by the North Carolina Education Research Data Center (NCERDC). These data provide comprehensive records of the universe of North Carolina public school students from 1993 through 2016. Key data elements include test scores, teacher and classroom assignments, demographics of students, parents, and teachers, and disciplinary and attendance records. We measure academic achievement based on end-of-year math and reading test scores in grades 3 through 8.

Our primary analyses focus on the impacts of elementary school teachers in grades 4 and 5. We chose elementary school teachers because middle school students in grades 6 through 8 typically see multiple teachers throughout the school day, complicating the measurement of effects for individual teachers (Jackson, 2014, 2018), although we explore the effects of middle school teachers in the appendix. Following Rothstein (2017), we use unique identifiers for the end-of-year test proctor to link elementary school students to their teachers. Lagged test scores are a key control variable in our value-added analysis (Chetty, Friedman, and Rockoff, 2016), meaning that this teacher-student match allows us to evaluate the quality of teachers in grades 4 and 5.

Our primary analyses focus on the impacts of elementary school teachers, in grades 4 and 5. Students in later grades typically see multiple teachers throughout the school day, complicating the measurement of effects for individual teachers (Jackson, 2014, 2018). We explore the effects of middle school teachers in the appendix. We link elementary school teachers to students using teachers that proctored the students' end-of-grade tests, similar to Rothstein (2017). These links exist from 1995 through 2011, giving us a 17 year panel on which to estimate teachers' test score impacts.

Our analysis also measures behavioral outcomes, most notably absences and suspensions. These outcomes exist for a shorter panel. Absences are available for all students beginning in 2004, and disciplinary records begin in 2001 for a fraction of the schools. Appendix 1.11 provides additional information regarding the construction of our panel.

Criminal Records

There are two sets of criminal records: charges and convictions. The criminal charges data come from the North Carolina District and Superior Courts, and the criminal convictions records are provided by the Department of Public Safety. The former records cover all cases disposed in North Carolina for offenses committed between 2005 and 2015. These data include detailed information on any criminal charges filed in the state that required the offender to appear in court, which constitutes the universe of charges all misdemeanor and felony offenses occurring within the state. In North Carolina, all individuals who are arrested must appear in court, so there is effectively no distinction between arrests and charges. Due to the relative frequency of criminal charges relative to convictions, we will primarily focus on these charge outcomes.

Conviction records are maintained by the Department of Public Safety. These data include all criminal convictions for offenses committed between 1970 through early 2017 that resulted in a mandatory supervision of the offender (probation or incarceration) as part of the offender's sentencing. In North Carolina, all felony convictions and severe misdemeanors have a supervision component to the sentencing, meaning that we observe the universe of serious criminal convictions in the state.

Sample Description

The NCERDC linked both sets of court records to the education data for all individuals born after 1989 for any adult (age 15 or above) criminal offenses occurring through the end of 2015. This implies we can measure criminal outcomes through age 25. The merge consisted of matches based on name, birth date, and last four digits of the social security number. Aggregate charge rates computed from our sample are similar to corresponding measures from official sources, which is reassuring and suggests that the match is accurate. More information about the match can be found in Appendix 1.11.

Table 1.1 displays summary statistics for elementary school students. (Descriptive statistics for middle schoolers appear in the appendix.) Roughly 25% of our sample are black and close to 50% are economically disadvantaged. Nearly a quarter of the sample has been charged with a criminal offense by age 20. The majority of these offenses are misdemeanors. However, over 5% of our sample has been charged with a felony. More than one-third have been charged with a misdemeanor, and over one-tenth of this cohort has a felony charge by the same age. Table 1.1 also reveals that offenders are much more likely to be black and male, less likely to have parents who attended college, more likely to have disciplinary problems in elementary school, and tend to have significantly lower test scores. The differences are much more pronounced for students with felony charges, with the differences even more stark for convicts (not shown).

1.3 Relationship Between Education and Criminality

We begin by using these linked education and crime data to establish three main sets of facts. First, early cognitive achievement measures are strongly associated with future criminality, even after controlling for a rich set of student characteristics. These observables include lagged test scores, indicating that test score gains are predictive of lower future criminality, in addition to levels. To the extent that gains are a better measure of the impacts of educational inputs, this fact suggests scope for educational quality to affect crime. Second, the correlation between test scores and future crime varies across demographic groups. Third, early behavioral markers are very strongly associated with future crime.

Fact I: Cognitive Achievement Measures Are Predictive of Future Crime

Early test scores are a powerful predictor of future crime. Figure 1.1 plots the average rate of criminal charges at age 20 in binned centiles of third-grade test scores. A one standard deviation increase in math scores in grade 3 is associated with a 4.8% drop in the likelihood of criminal charges. Similarly, a standard deviation increase in reading scores is associated with a 5.4% decrease in criminality.

These relationships continue to hold after adjusting for observed student characteristics. Panel A of Figure 1.2 plots partial coefficients from regressions of crime on test scores by grade, controlling for gender, race, family characteristics (parental education level, native language, and economic disadvantage status), behavioral measures (disciplinary record, attendance record, and if they repeated the given grade), school, and year. Two notable patterns appear here. First, controlling for observables cuts the relationship between crime and academic achievement roughly in half: a one standard deviation increase in third grade math (reading) scores is associated with a 2.26 (2.83) percentage point decrease in the probability of a criminal charge by age 20. These are approximately 10% of the mean charge rate. Second, the association between test scores and future crime is larger in magnitude for older students, particularly for math scores. The decrease in crime associated with a standard deviation increase in eighth grade test scores is 3.99 percentage points, nearly double that of a similar change in third grade test scores.[4]

Studies of teacher and school value-added suggest that test score gains are a better measure of educational input quality than test score levels. Panel B of Figure 1.2 plots the coefficient of a regression of criminal charges at age 20 on test score gains, conditional on controls. We still see large, significant associations between test score gains at all grades. This fact suggest that teachers and schools that improve test scores may also reduce criminality. We will address this concept more rigorously in the next section.

Fact II: Test Score-Crime Relationship Varies Across Demographic Groups

Figure 1.3, panel A shows the difference between associations of test scores and criminal charge rates for boys versus girls.

This figure reveals a much stronger relationship for boys—boys with a standard deviation higher math score are 4 percentage points less likely to have a criminal charge by age 20, approximately 50% higher likelihood than for girls. While this gap appears large, the proportional impacts for both groups are approximately equal; in our sample, 29% of boys have a criminal charge on their record by this age, as compared to 17% of girls. This ex-ante gap in criminal charge rates is mostly, but not entirely, covered by the relative baseline offending rates.

We see a gap in the absolute association rates across races, although the proportional effects are approximately equal. Panel B in Figure 1.3 shows the difference in associations in arrest rates by age 20 and math scores for black students and white students. We see that the association between test scores and crime is larger in magnitude for blacks than it is whites. Again, this gap appears to be partially driven by the differences in average offending

[4] Moreover, when we run the horserace include multiple years of test scores, we find that the association loads onto the most recent test score. Interestingly enough, we run the same exercise with in which we put *all* available test scores in this regression. We find that the association loads most strongly onto the eighth grade test score, indicating that the most recent measure of cognitive ability is the most salient.

rates – 30% of black students have been arrested by that age, as compared to 20% of white students, implying that this gap is likely due to the baseline differences across racial groups, although the relative percentage point gap is smaller across races.

In contrast to the previous groups, the test score-crime relationship appears similar for groups of different economic status backgrounds, although there is a gap in relative terms when accounting for baseline offending rates. Panel C in Figure 1.3 shows the relative rates for students from economically disadvantaged backgrounds relative to those who are not disadvantaged. The rates are practically indistinguishable. However, students from disadvantaged backgrounds are about 50% more likely to have been charged with a crime by age 20 (29% vs 19%, respectively). In essence, this means that the magnitude is stronger for non-disadvantaged students in terms of the gradient relative to baseline offending rates, suggesting that associations between test scores and crime matter more for people for students from more privileged backgrounds.

Fact III: Early Behavior Markers and Future Criminal Behavioral Are Strongly Correlated

Figure 1.4 shows the link between our rough behavioral measures in early grades, namely disciplinary records and attendance records, and future crime. Early suspensions are a strong predictor of future criminal behavior; 50% of students with a suspension in elementary school have been charged with an offense by age 20. Panel A shows that elementary school kids are 28 percentage points more likely to commit a crime than those who have not been suspended; we are able to account for about half of this gap using descriptives about the student. Suspensions are rare in elementary school (about 5% of students are suspended), but this indicates that suspensions are an indication for at-risk students.

Similar patterns arise for attendance records. In order to match our current analysis with later analysis, we choose to plot the relationship between $\log(\text{Absences} + 1)$.[5] Similarly to the discipline records, higher truancy rates are strongly associated with higher future crime rates. Additionally, unlike suspensions, however, there is a clear break in relationships between truancy and its association in middle school vs elementary school, with a much larger effect in later adolescence.

Additional Heterogeneity and Summary

In the appendix, we explore additional statistical relationships between crime rates and observable outcomes in elementary and middle school. Similar patterns exist associating test scores and criminal behavior when considering the severity of the criminal infraction and convictions versus charges. We also see similar patterns when varying the age of the

[5] We could use integer valued attendances or more hyperbolic inverse signs to be able to use percentage interpretations of a variable that can be integer-valued. We chose to follow the literature (Jackson, 2018), although the results are qualitatively similar, regardless of the specification.

offense, although we see slightly stronger relationships for criminal behavior at later ages, indicating that test scores are not merely a signal of the timing of your first offense, but also the likelihood of committing any crime.

We also considered the relationships between test scores and types of criminal infraction. Approximately half of our criminal charge records are classified into several broad categories: assaults and violent crimes, property crimes (e.g. burglary), and drug offenses. In general, higher test scores are associated with lower crime rates for all three classifications of charge, and similar heterogeneity patterns exist for these outcomes as the patterns documented in Section 1.3. One notable exception is the test-score/crime relationship for drug offenses - contrary to what was shown earlier, the association between test scores and is *weaker* for economically disadvantaged and black students.

In summary, there is a clear link between early cognitive achievement tests, behavior in elementary school, and crime. If we are willing to assume that (1) this relationship is not entirely due to unobservable characteristics of the student and (2) teachers impacts on cognition and behavior have long-lasting impacts, then the above descriptive relationships suggest that teachers who impact short-run outcomes such as test scores and behavior may be able to improve long-run crime outcomes. We will proceed by estimating teacher quality directly to test for its relationship to students' future criminal behavior.

1.4 Estimation of Teacher Effects

Univariate Value-Added Model

Following the prior literature on teacher value-added (Chetty et al., 2014a; Jackson, 2018), we begin our analysis of teacher effectiveness with an empirical specification that relates academic achievement to student characteristics and teacher quality as follows:

$$A_{it} = X_{it}'\beta + \alpha_{j(i,t)} + \epsilon_{it} \tag{1.1}$$

Here, A_{it} is a measure of achievement, such as a test score, of student i in year t and $\alpha_{j(i,t)}$ is the casual impact of teacher j on i's achievement in year t. The vector X_{it} includes observed characteristics of the student. Importantly, this set of controls includes the lagged value of the outcome A_{it-1}, allowing us to interpret the teacher effects α_j as differences in test scores for students with similar past achievement. The control vector also includes cubics in lagged math and reading test scores, gender, race, parental education level, economic disadvantage status, limited English proficiency, special education status, grade repetition, lagged suspension and log absences, class size, and year fixed effects, as well as school-level averages of all the above control variables.

A causal interpretation of the α_j parameters requires a "selection on observables" assumption:

$$\mathbb{E}[\epsilon_{it}|X_{it}, j(i,t)] = 0 \tag{1.2}$$

This assumption requires teacher assignments to be independent of students' potential achievement conditional on lagged test scores and the other variables in X. Chetty et al. (2014a) argue that the set of controls used here is sufficient to isolate causal impacts of teachers on test scores. Equation (1.1) also imposes a constant effect of a given teacher across years. This parsimonious specification allows us to produce a precise single measure of value-added for each teacher, which may be interpreted as a weighted average effect across years if teacher effects "drift" over time (Angrist and Pischke, 2009; Chetty et al., 2014b).Even assuming constant effects over time, our estimates of the α_js may be noisy. As is standard in studies of teacher value-added, we will use empirical Bayes (EB) posterior estimates of our teacher effects to "shrink" noisy estimates of teacher effects and reduce mean squared error. The shrinkage is based on the specification:

$$\widehat{\alpha}_j = \alpha_j + e_j \tag{1.3}$$

where $\hat{\alpha}$ is an estimate from OLS estimation of (1) and e_j is estimation error. Studies of teacher value-added typically model α_j as normally distributed with constant mean and variance conditional on X:

$$\alpha_j | \boldsymbol{X}_j \sim N(\alpha_0, \sigma_\alpha^2) \tag{1.4}$$

We extend this approach to allow the distribution of α_j to depend on X for two reasons. First, the assumption that X is independent of alpha is empirically falsifiable: about 25 percent of the variation in α_j is explained by class means of X. Shrinkage measures that do not account for this dependence will be less accurate. Second, shrinking conditional on X is useful for our analysis of longer-run outcomes, as described further below.

Our conditional shrinkage is based on the model:

$$\alpha_j | \boldsymbol{X}_j \sim N(\bar{X}_j' \gamma, \sigma_\alpha^2) \tag{1.5}$$

where $\boldsymbol{X}_j$ is the matrix of characteristics for all students in teacher j's class and $\bar{X}_j$ is the class mean of these characteristics. This model allows higher value-added teachers to be assigned to classes with systematically different observables, as in "correlated random effects" panel data models (Chamberlain, 1980).

The minimum MSE prediction of teacher j's effectiveness is then:

$$\alpha_j^* \equiv \mathbb{E}[\alpha_j | \hat{\alpha}_j, \boldsymbol{X}_j] = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + Var(e_j)} \hat{\alpha}_j + \left(1 - \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + Var(e_j)}\right) \bar{X}_j' \gamma \tag{1.6}$$

Rather than shrinking the unbiased teacher effect $\hat{\alpha}_j$ towards the overall mean, this prediction shrinks $\hat{\alpha}_j$ towards a conditional mean that depends on $\bar{X}_j$. We estimate γ and σ_α^2 using maximum likelihood, and calculate $Var(e_j)$ as the squared standard error of $\hat{\alpha}_j$.

Long Run Effects

The goal of our analysis is to relate estimates of teacher effects on short-run outcomes to effects on crime. We investigate this relationship with the following specification:

$$Y_{it} = X'_{it}\eta + \delta\alpha_{j(i,t)} + \nu_i \tag{1.7}$$

Here, Y_{it} is a criminal outcome, and $\alpha_{j(i,t)}$ is the causal effect of teacher j on a short-run outcome (either cognitive or non-cognitive). This specification parallels Chetty et al.'s (2014b) analysis of the relationship between teacher value-added and adult earnings. The parameter δ measures the extent to which teachers that improve short-run outcomes also reduce crime. Similar to our assumption for unbiased teacher effects α_j in (1.1), identification of δ requires student unobservables in the crime equation to be independent of teacher assignment conditional on observables.

We do not observe the true causal effect α_j. We instead observe our value-added estimates α_j^*. In our case, we will run the following regression:

$$Y_{it} = X'_{it}\tilde{\eta} + \tilde{\delta}\alpha^{*z}_{j(i,t)} + \tilde{\nu}_{it} \tag{1.8}$$

Here, $\alpha^{*z}_{j(i,t)}$ is the standardized value of our estimated teacher effects from (1.6). The parameter of interest, $\hat{\tilde{\delta}}$, is the change in criminal likelihood due to a standard deviation change in teacher quality.

An alternative approach to estimating teacher impacts on crime is to fit the model in (1.1) with crime on the left-hand side. Chetty et al. (2014b) argue that the availability of a lagged outcome control is necessary for selection on observables to hold, so that OLS estimates for longer-run outcomes like crime may be biased even while estimates for short-run outcomes are unbiased. In this case, Equation (1.8) will give us the relationship between crime effects and short-run value-added even if unbiased estimates of crime effects for individual teachers are unavailable. We therefore report estimates of (1.1) for crime in addition to estimates of (1.8), but interpret the former with caution.

Identification of Teacher Impacts

Our estimation proceeds in three steps: estimation of short-run teacher impacts, shrinkage to reduce mean squared error, and estimation of the relationship between crime and short-run impacts. Each of these three stages requires specific assumptions in order to make inference. We discuss these assumptions below, and when applicable, describe our tests for such assumptions.

First, the key assumption underlying our value-added estimation strategy is selection on observables. If students are assigned to teachers based on characteristics that we cannot observe, our estimates of teacher impacts will be biased. A large portion of the literature is comfortable with selection on observables for test score outcomes,[6] and recent work in the non-cognitive VA literature similarly validates this assumption for behavioral outcomes (Petek and Pope, 2016; Jackson, 2018).

To justify this assumption in our context, we follow Chetty et al. (2014a) and use teacher mobility "experiments" to test for bias in our estimates of teacher value-added. This specification check is based on estimation of the equation:

$$\Delta A_{sgt} = \lambda_1 \Delta Q_{sgt} + \lambda_2 \Delta \chi_{sgt} \tag{1.9}$$

All quantities in the above regression are aggregated to the school-cohort level for a given school s, grade g, and year t. Q is the student-weighted average of teacher VA measures across cohorts, A_{sgt} is the school-cohort average test score, and χ_{sgt} is a school-cohort level average of controls. If our VA estimates are accurate measures of teacher quality, then aggregate changes in our teacher quality measure at the school should on average be equal to the aggregate change in the achievement measures A after controlling for other observables. This will hold for when $\lambda_1 = 1$; under this condition, our teacher effect estimates are "forecast unbiased", meaning that they on average capture the true effects of teachers on A. Similar tests in other studies estimate this model using teachers who change schools or grades as a source of exogenous variation in teacher quality across schools (Chetty et al., 2014a).

Estimates of λ_1 for all VA measures are shown in the appendix. For the majority of the different achievement measures, the estimates of λ_1 are indistinguishable from 1. In our context, however, the relationship may be mechanical. In our estimation of (1.9), estimates of λ_1 are identified from both changes in teacher composition and changes in quantity of student exposure to different teachers within the same school. In the future, we plan to identify aggregate changes in outcomes at the cohort level using only exogenous teacher switches to further justify selection on observables. For the time being, we will rely on these imperfect, but suggestive estimates, and will refer to previous studies suggesting that selection on observables holds for cognitive (Kane and Staiger, 2008; Chetty et al., 2014a) and non-cognitive (Petek and Pope, 2016; Jackson, 2018) VA models.

The second important assumption is made in the EB shrinkage procedure, which assumes a normal prior on the teacher quality distribution. Misspecification could lead to incorrect inference. This threat to validity is particularly concerning for our discrete outcomes, such as our disciplinary records. We therefore test a variety of shrinkage procedures, including an unconditional shrinkage, and no shrinkage. We elaborate on these estimates in Section 1.5. Changes in our shrinkage procedure do not qualitatively affect the results.

Third, our final specification (1.8) requires us to assume that teacher quality is not related to unobservable determinants of the students' criminal behavior. Additionally, unlike

[6] Notably, Rothstein (2010) finds measures of bias by finding association between future teacher quality and prior test score gains. Chetty et al. (2014a) develop their own test, comparing effects of teachers of different quality switching schools and grades, and find that the aggregate school-cohort level changes in outcomes due to these shocks match those predicted by the VA measures. There is still some debate on if this test is indeed exogenous - Rothstein (2017) has raised questions about this; Bacher-Hicks, Kane, and Staiger (2014) and Chetty, Friedman, and Rockoff (2017a) address these concerns.

equation (1.7), our measure of teacher quality α_j^{*z} is estimated, which in small samples may be mechanically correlated with $\tilde{\nu}$, biasing our estimate of $\tilde{\delta}$ (Jacob, Lefgren, and Sims, 2010). We avoid this mechanical correlation by using "jackknifed" estimates of α_j^{*z}, done by estimating (1.1) separately for all t using all years $s \neq t$. By implementing this "leave-year-out" procedure, the resulting teacher effects $\alpha_{j(i,t)}^{*z}$ are estimated without using i's cohort, that is, by estimating $\alpha_{j(i,t)}^{*z}$ using all students k in years $s \neq t$, eliminating the mechanical correlation to i's error term.

Relationship to 2SLS

It is useful to link our approach to estimating equation (1.8) to issues that arise in the econometrics of instrumental variables models. In estimating (1.8) we seek to recover the relationship between teachers' effects on crime and their effects on test scores. One can view this as an instrumental variables problem in which teacher indicators are used as instruments for test scores A_{it} in an equation for crime Y_{it}, in the following specification, as follows:

$$\begin{aligned} Y_{it} &= X_{it}'\eta + \delta A_{it} + \nu_{it} \\ A_{it} &= X_{it}'\beta + \alpha_{j(i,t)} + \epsilon_{it} \end{aligned}$$

In this view, the first stage equation of this system corresponds to (1.1), and δ from the second stage represents the same parameter of interest as δ from (1.7). The "leave-year-out" procedure described earlier in this section is useful because it reduces many/weak instrument bias that would arise for 2SLS due to small sample sizes for each individual teacher. The parameter of interest, δ, captures the causal impact of a change in A_{it} due to variation in teacher quality on Y_{it}.

In Appendix 1.12, we show the direct connection between 2SLS and VA regression frameworks explicitly. This framing of long-run teacher impacts as an instrumental variables problem allows us to think about the issues that arise in estimating long-run teacher effects. There are three technical issues here. First is the exclusion restriction made in IV. To gain identification in this framework, we are essentially assuming that the only channel through which teachers impact their students are through their direct influence on the ability measure A_{it}. For the time being, let's suppose that this measure is test scores. If teachers are influencing skills other than cognitive measures, than this will almost certainly fail.

The second technical issue is the motivation for the jackknife estimator, as there may be a problem with many weak instruments here (Bound, Jaeger, and Baker, 1995). We see that by adding numerous teacher dummies, we may be overfitting our first stage equation (1.1), mechanically inducing correlation between our estimated fitted values $\hat{A}_i$ from the first stage and second stage error due to the correlation in unobservables for our equations for Y_{it} and A_{it}. In the appendix, we also show that, given our estimates, a jackknifed IV estimator will approximately recover δ.[7] The rationale of this result parallels that of the

[7] A minor caveat here: the proof shows this for "leave-person-out" averages, whereas we use "leave-year-out" averages, although the result can easily be extended to incorporate more general averages.

2SLS framework, which shows jackknifed fitted values are uncorrelated with second-stage regression errors (Angrist, Imbens, and Krueger, 1999). This result validates our approach to accurately recover long-run impacts of teachers.

Third, this 2SLS approach reveals that standard errors generated by naive estimation of (1.8) are likely to be conservative. This is a special case of the fact that standard errors from "manual 2SLS" procedures are incorrect, and generally too large except under extreme forms of endogeneity (Angrist and Pischke, 2009). The estimates when following our three-step procedure are still very precise; we do not implement any further corrections to our standard errors beyond clustering at the school-cohort level.

1.5 Results from Univariate Model

Estimates of Teacher Quality

Our main estimates focus on the impacts of elementary school teachers. To estimate VA for achievement measure A_{it}, we restrict our estimation sample to students with observed lagged values of the outcome who have been matched to teachers observed for multiple years with at least 25 students. We estimate teacher value-added using Equation (1.1) for three types of short-run measures A_{it}. First, we report VA estimates for reading and math test scores, which we will refer to as cognitive VA.

Second, we construct estimates of non-cognitive VA using suspensions and absences. Prior studies have used individual behavioral measures (Gershenson, 2016; Holt and Gershenson, 2017) and weighted averages of multiple behavioral outcomes and letter grades (Petek and Pope, 2016; Jackson, 2018) to measure students' non-cognitive achievement. We chose the former approach both for clarity in interpretation of our estimates, and because in North Carolina, the "grades" reported for elementary school students are effectively noisier test score measures. However, unlike test scores, these behavioral actions can be directly influenced by the teacher. In particular, effects on contemporaneous suspensions capture both students' changes in juvenile behavior and teachers' differential propensity to punish their students. In order to capture just the student component, we also use future suspensions and absences in grade 6 to evaluate achievement in elementary school (Petek and Pope, 2016).

Third, we construct estimates of "crime value-added", in which we estimate (1.1) using age 20 crime outcomes for A_{it}. As discussed in 1.4, we do not include a lagged value of our outcome in our controls for this estimate, potentially generating biased estimates. Assuming positive selection of students of high ex-ante crime propensity to high "crime VA" teachers, the magnitude of our effects will be biased upward. Thus, these "crime VA" estimates will bound the extent to which teacher employment policies can be used to change crime.

Table 1.2 reports our estimates of σ_α for each A_{it}. As a reference, Column (1) estimates a model without controls to show the variation in test scores without controlling for observables. Column (2) reports the variation in these VA measures for the given outcomes that we use in our paper, after controlling for our robust set of demographics. We have show the distribution of teacher impacts on two cognitive measures (math and reading tests), four non-cognitive measures (contemporary and future suspensions and absences), " crime VA" for criminal charges by age 20.

Teacher quality varies substantially for each outcome. Similar to other previous work, we estimate the standard deviation of the teacher impact α_j to be 0.2 test score standard deviations in math and 0.1 test score standard deviations in reading. In other words, a teacher who is one standard deviation higher in the distribution of math (reading) value-added improves math and reading test scores by almost 0.2 and 0.1 student test score standard deviations, respectively. Teachers also affect non-cognitive outcomes: we find that the standard deviation of α_j for grade 6 suspensions is 0.034, meaning that a one standard deviation improvement in elementary school teacher quality reduces the probability of suspension in grade 6 by 3.4 percentage points. Our preferred non-cognitive models use these grade 6 outcomes; we also report estimates for contemporaneous suspensions and absences in elementary school, but since these are under direct control of the elementary school teacher they may conflate effects on students' skills with a direct influence on the outcome (Petek and Pope, 2016).

Long Run Estimates

Table 1.3 summarizes our estimates of Equation (1.8). These estimates come from specifications that treat each student-year in elementary school as an observation. Our outcome of interest is whether or not the student have committed a crime by age 20, measured by severity of crime (any crime versus felony) and judicial status of the criminal infraction (criminal charge versus conviction). We interpret the coefficients as the partial effect of a one standard deviation increase in teacher quality on either the cognitive dimension or the non-cognitive dimension.

Relationships between test score value-added and future crime are generally weak. Specifically, a one standard-deviation increase in teacher quality is associated with a 0.1 percentage point decrease in charge rates by age 20, or about 0.4% of the mean conviction rate at age 20. Estimated effects for reading on felony charges and conviction outcomes are similar. Math VA has a statistically insignificant impact on criminal charges, although we do see similar effect sizes on criminal convictions.

In constrast, teachers that improve non-cognitive outcomes in the short run reduce future crime. This can be seen in the remaining rows of Table 1.3, which report relationships between criminal outcomes and estimated teacher effects on absences and suspensions. Note that the sign of the coefficients on our behavioral measures is now positive, indicating that elementary school teachers that increase these measures (i.e. cause increases in middle school absences or suspensions) generate worse crime outcomes. We have reported teacher's suspension and attendance VA affects long-run outcomes. On average, teachers that increase the likelihood of a grade 6 suspension by 1 standard deviation also boost the likelihood of criminal charges by age 20 by 0.42 percentage points, or 1.8% of the mean charge rate. A one standard deviation increase in a teacher's effect on the absence rate leads grade 6 by one standard deviation lead to a 0.48 percentage point increase in criminality (2.1% of the mean charge rate). These estimates are large, precise, and qualitatively consistent across all of our crime measures.

To gauge what types of crimes may be influenced by better teachers, we also estimated the same regressions focusing on felony charges by age 20.[8] The marginal impact of a standard deviation change in non-cognitive VA is about half of that for overall charges. Recalling that our felony charge rate is about a quarter of our overall charge rate, this suggests that good teachers are disproportionately better at reducing serious offenses for students. We see similar patterns for effects of non-cognitive VA on conviction outcomes. We will explore the implications of these results in section 1.7.

In addition to estimating differences in the type of crime outcome, we test for heterogeneity in the effects of teacher VA for different demographic groups. In the appendix, Table 1.9 estimates (1.8) separately by race. In general, effects of high non-cognitive VA are larger for black students relative to white students, particularly for the more serious felony and conviction outcomes. Similarly, Table 1.10 displays estimates (1.8) separately by socioeconomic status. The effects of high non-cognitive VA for students from economically disadvantaged backgrounds are approximately double the effects of individuals from more privileged backgrounds. These results suggest that teachers who are good at imparting non-cognitive skills onto students may be better at reducing crimes for underprivileged populations.

The previous results estimate the effect of teacher quality on students' likelihood of their first offense occurring before age 20. Table 1.11 also shows effects at various ages of first offense. At all ages, the effects of cognitive VA on criminal activity are small, but the effects of non-cognitive VA are large and robust. Moreover, the magnitudes of the effects increase with the age of the first offense, indicating that teachers are affecting crime rates over time. Note that while the criminal outcomes are monotonically increasing by age, (if one were charged with your first offense before age 20, they would also be charged before age 21), that does not necessarily mean that the relationship between crime and teacher quality is monotonically increasing. For example, it is possible that teachers of varying quality simply shift criminal activity to different time horizons, but the effect on aggregate lifetime criminal behavior does not vary with teacher quality. Given that the magnitudes of the effects increase with the age of the first offense, this hypothesis seems unlikely. Teachers with high VA are affecting their students' future criminal behavior.

Alternative Specifications

To verify our central findings, we estimate the effects of teacher quality on crime using a variety of specifications. These estimates can be found in the appendix. The general conclusions of these alternative specifications match those from the main results. These estimates provide suggestive evidence that teachers can impact crime outcomes through

[8] If we restrict our outcome to less severe offenses, i.e. misdemeanors, we see effects that are almost exactly the same as the effects on overall charge rates. This is likely because over 95% of people who have been charged with a crime by age 20 have been charged with a misdemeanor; rarely do we find individuals with a felony charge and no misdemeanor charge.

a cognitive dimension, although these effects are small and generally inconclusive. More certainly, there is a non-cognitive component of teacher impacts that have lasting results on their students' propensity for future criminal behavior.

Our preferred specification pools teacher effects across grades and treats each student-year as a separate observation. This pooled specification utilizes all available data to precisely estimate the impact of teacher quality. However, this specification fails to account for correlation in VA of a given student's teachers in consecutive years. This phenomenon results from correlated estimation errors of adjacent teachers due to shocks in achievement measures (Rothstein, 2010, 2017). To address this problem, we estimate several different alternative specifications that do not have multiple appearances of similar students. Table 1.12 displays estimates of (1.8) run separately by grade. In this case, the test score VA is weakly significant for the grade 4 teachers and very small and insignificant for the grade 5 teachers. However, for both grades, the impacts of high non-cognitive VA has larger, significant impacts on all crime outcomes.

Another way to address correlated VA across years in the pooled regression is to aggregate the teachers' VA for each student across grades to test how the total impact of a student's fourth and fifth grade teacher impacts her future crime. Table 1.13 reports the estimates of cumulative and averaged VA in elementary school on crime.[9] The estimated effects are smaller in magnitude and noisier, particularly for the test score outcomes. Again, teachers' non-cognitive VA, particularly when measured with future suspensions, has the largest impact on future crime.

To address the possible misspecification in our shrinkage, we estimate (1.8) using two alternative VA measures: teacher effects using the unconditional shrinkage procedure described in (1.4) and using unshrunken teacher effects from (1.1). These estimates have been reported in Table 1.14. Results from Section 1.12 indicate that the shrinkage procedure affects the distribution of VA, but the teacher effects on crime should not change substantially under different shrinkage methods. Indeed, while the shrinkage procedure affects the distribution of teacher effects on short-run outcomes, we find extremely similar impacts of teacher quality on crime.

Additionally, we estimate similar crime effects for middle school teachers. Table 1.15 displays these estimates. As mentioned before, these results are harder to interpret than elementary school teacher effects, these crime effects are partial holding all other instruction constant. For these middle school teachers, there is still a strong, significant impacts of non-cognitive VA on future criminality. There is also some evidence that better reading VA teachers have strong impacts on crime, although the impacts of math VA remain small and insignificant.

The final two appendix tables address biases in crime estimates due to migration. We only observe criminal acts committed in North Carolina. If out-of-state crime by North Carolinian students is substantial and related to teacher assignment, then the previous estimates would

[9] These estimates differ from one another because we can only match approximately 2/3 of students to teachers in our sample, which limits the number of students in the cumulative VA sample.

mischaracterize the relationship between VA and crime. We address this by re-estimating (1.8) for students that appear in our education records in grades 9 and 12. For this sample, the estimates for early crime outcomes are unaffected by migration these students must still reside within North Carolina. Table 1.16 reports estimates of (1.8) for these populations on crime outcomes at age 18, which necessarily appear in the crime data for the grade 12 sample. The results are comparable to the main results. Table 1.17 reports the same estimates for age 20 crime outcomes, and yields similar conclusions. We find no evidence of migration significantly affecting our estimates.

In summary, we find that teachers affect their short-run test scores and behavioral outcomes, and that the latter effects of teachers lead to long-run changes in criminal behavior. We now turn to a multivariate random effects model to learn more about the joint distribution of teacher effects.

1.6 Multivariate Random Coefficients Approach

Model and Estimation

The estimates in Table 1.3 capture bivariate relationships between crime and a given measure of value-added. Teacher effects on cognitive and non-cognitive dimensions may be correlated, however. Figure 1.11 in the appendix displays binned scatter plots of teachers' cognitive and non-cognitive VA. This visual evidence is not proof that these effects are uncorrelated. These teacher effects are estimated with error, meaning that any existing correlation would be attenuated due to noise. Moreover, these cognitive and non-cognitive VA measures are estimated independently of one another. To determine the joint distribution of teacher effects along cognitive and non-cognitive dimensions, estimates must consider the joint distribution of the measures of cognitive and non-cognitive measures, particularly the direct correlation of teachers' effects on multiple outcomes, and the correlation in unobservable characteristics of the student.

To more fully explore the correlation structure between measures of value-added and their link to future crime, we now extend the model to a multivariate random effects framework that jointly estimates teacher effects on each outcome, allowing us to recover partial effects of each dimension on future crime. This approach will also take seriously the binary nature of the non-cognitive outcomes we consider, which may lead to misspecification issues in OLS estimation.

The multivariate random effects model is based on the following specification:

$$A_i^k = X_i' \beta^k + \alpha_j^k + \epsilon_i^k \tag{1.10}$$

This approach follows Broatch and Lohr (2012) to estimate multiple dimensions of teacher effects. This specification is similar to Equation (1.1), with k indexing the short-run outcomes. The multivariate set-up allows for flexible correlation between the teacher effects α_js and the unobservables of the student ϵ_i.

In our case, we use one cognitive and one non-cognitive measure of teacher effects. In our implementation, we use reading scores and future suspensions, respectively. As our preferred non-cognitive measure is binary, we choose to use a probit specification for our secondary outcome. Specifically, we fit the following model:

$$\begin{aligned} A_{it}^1 &= X_i'\beta^1 + \alpha_{j(i,t)}^1 + \epsilon_i^1 \\ A_{it}^{2*} &= X_i'\beta^2 + \alpha_{j(i,t)}^2 + \epsilon_i^2 \\ A_{it}^2 &= \mathbb{1}\{A_{it}^{2*} > 0\} \\ \alpha_j|\boldsymbol{X} &= (\alpha_j^1, \alpha_j^2)' \sim N\left(\bar{X}_j'\gamma, \Sigma_\alpha\right) \\ \epsilon_i|\boldsymbol{X} &= (\epsilon_i^1, \epsilon_i^2)' \sim N\left(0, \begin{bmatrix} \sigma_e^2 & \rho_\epsilon\sigma_e \\ \rho_\epsilon\sigma_e & 1 \end{bmatrix}\right) \end{aligned} \tag{1.11}$$

Here, A_{it}^1 is the reading test score and A_{it}^2 is the suspension occurrence in middle school. We assume normal distributions for the teacher impacts α_j and the student unobservables ϵ_i. The covariance matrix $\Sigma_\alpha = \begin{pmatrix} \sigma_{\alpha,1}^2 & \rho_\alpha\sigma_{\alpha,1}\sigma_{\alpha,2} \\ \rho_\alpha\sigma_{\alpha,1}\sigma_{\alpha,2} & \sigma_{\alpha,2}^2 \end{pmatrix}$ describes the variation in effectiveness across teachers on each skill dimension as well as the correlation between teachers' effects on cognitive and non-cognitive skills. The parameters σ_e and ρ_ϵ govern the distribution of unobservables at the student level. As in our univariate estimation, this model allows for a correlated random effects structure in which the mean of the distribution of teacher effectiveness is depends on student characteristics. We estimate this model using simulated maximum likelihood.

Posterior Estimation and Teacher Impacts

We use estimates of (1.11) to form EB predictions of teacher quality that account for both cognitive and non-cognitive outcomes jointly. In this multivariate approach, however, there is no closed-form estimate of the EB posterior mean. We therefore use the posterior mode instead, following Angrist, Hull, Pathak, and Walters (2017). Formally, we define our posterior estimates $\tilde{\alpha}_j$ as follows:

$$\tilde{\alpha}_j = \underset{\alpha_j}{\arg\max}\ f(\alpha_j|\beta^1, \beta^2, \Sigma_\alpha, \sigma_\epsilon^2, \rho_\epsilon, \boldsymbol{A}_j, \boldsymbol{X}) \tag{1.12}$$

Here, f is the density of α_j, conditional on the short-run outcomes of teacher j's students ($\boldsymbol{A}_j$), all observable student characteristics $\boldsymbol{X}$, and the parameters in the model (1.11). The posterior mode $\tilde{\alpha}_j$ represents js most probable teacher quality given the observed characteristics of her classroom.

The resulting non-cognitive teacher effect, $\tilde{\alpha}_j^2$, is a measure of the teacher's effect on the latent likelihood of suspensions. To assist with interpretation, we instead use the following measure for non-cognitive VA in our estimates of crime outcomes:

$$g(\tilde{\alpha}_j^2|\bar{X}) = \Phi(\bar{X}'\beta^2 + \alpha_j^2) - \Phi(\bar{X}'\beta^2 + \mu_\alpha) \tag{1.13}$$

Here, $g(\tilde{\alpha}_j^2|\bar{X})$ represents the change in future suspension probability for the mean student who moves from a teacher of mean latent quality $\mu_\alpha = \bar{X}'\gamma$ to a teacher with latent quality $\tilde{\alpha}_j^2$. Our estimates focus on this measure of teacher quality, as the probabilities are directly interpretable, and are in the same units as the univariate effects. Unlike the univariate model, teacher quality changes on the probability of future suspensions differ by student observables X, so we focus on the impact for the average student.

Estimates from the multivariate teaching model are very similar to those from the univariate specifications decribed in Section 1.4.[10] Table 1.4 reports estimated parameters in our correlated random effects model, and compares them to the univariate estimates based on the methods described in Section 1.4. Implied marginal effects on the outcomes of improved teacher quality based on both methods are comparable. The correlation between estimated posteriors from the multivariate and univariate estimates is 0.8, indicating that while the two approaches yield similar results they do not generate identical predictions.

The estimated correlation between cognitive and non-cognitive value-added (α_j^1 and α_j^2) is 0.027, statistically indistinguishable from zero. This implies teachers' impacts on students' future discipline and contemporary test scores are unrelated. This zero correlation between teacher effects is in contrast to a strong correlation at the student level. The correlation between our unobserved errors in reading scores and future suspensions is a small but statistically significant -0.063, indicating that students who are unobservably better at taking tests are also more likely to be suspended in the future.

Long-Run Effects on Crime

Given our new sharpened estimates of the vector of teacher quality, we can then estimate long run impacts. We estimate the following equation:

$$Y[it] = X_{it}'\eta + \delta_1 \tilde{\alpha}_{j(i,t)}^1 + \delta_2 \times g(\tilde{\alpha}_{j(i,t)}^2|\bar{X}) + \nu_i \tag{1.14}$$

This specification allows us to measure partial relationships between crime effects and cognitive and non-cognitive effects. Similar to estimation of equation (1.8), estimates of cognitive ($\tilde{\alpha}_{j(i,t)}^1$) and non-cognitive ($g(\tilde{\alpha}_{j(i,t)}^2|\bar{X})$) ability are standardized to have mean zero and standard deviation one. This specification assumes additive separability of teachers' different short-run effects on their students' criminal behavior, allowing us to interpret δ_1 (δ_2) as the partial effect of a one standard deviation increase in cognitive (non-cognitive) VA on future crime probability holding other dimensions of teacher quality constant.

Table 1.5 displays estimates of equation (1.14). Column (3) reports estimates of the full specification for our crime outcomes at age 20. The results show that teacher quality impacts on crime arise entirely through a non-cognitive channel. The effects of reading test scores are small and insignificant, whereas the non-cognitive effects are large, with a

[10] Similar to Equation (1.13), we report $\tilde{\sigma}_{\alpha,2} = \Phi\left(\bar{X}'\beta^2 + \sigma_{\alpha,2} + \mu_\alpha\right) - \Phi\left(\bar{X}'\beta^2 + \mu_\alpha\right)$ in row 2, column 4. This is interpreted as the change in probability of a future suspension for the average student when moving from a teacher of average quality to a teacher with a standard deviation higher in latent suspension VA.

standard deviation increase in $(g(\tilde{\alpha}^2_{j(i)}|\bar{X}))$ leading to a 0.6 percentage point decrease in criminal charges by age 20. Relative to the mean occurrence, the effects are even larger for felony charges and convictions. Columns (1) and (2) also report these univariate effects by only including the cognitive and non-cognitive VA estimates in Equation (1.14), respectively. These point estimates are almost identical as the estimates from the full model, as expected since these teacher effects are orthogonal.

We also estimated (1.14), replacing the multivariate VA $\tilde{\alpha}^1_{j(i)}$ and $g(\tilde{\alpha}^2_{j(i)}|\bar{X})$ with the corresponding VA estimates from our univariate model. These estimates are reported in Column (6), with (4) and (5) showing the univariate effects.[11] The results are almost identical to the estimates using multivariate VA.

In summary, we used a correlated random effects framework to account for multidimensional teacher effects. The results indicate that teachers' short-run effects on cognitive and non-cognitive skills are orthogonal. Moreover, teachers' effects on their students' crime is found to be entirely driven by their non-cognitive effects on students. These results indicate that hiring policies using test score VA will fail to even partially compensate teachers' development of their students' non-cognitive skills (Neal, 2011). As these skills generate potentially large welfare gains through reduced crime, such test-score based hiring policies may be suboptimal. We will explore this notion in the next section, where we use our estimated teacher impacts to quantify how teacher hiring policies can be used to reduce crime.

1.7 Policy Consequences

Our results have implications for the design of teacher personnel policies that use measures of teaching effectiveness as inputs. As teacher quality is not well predicted by ex-ante observable characteristics of the teacher (Rivkin, Hanushek, and Kain, 2005), but many districts use measures of test score value-added for firing and promotion decisions (Fryer, 2013; Podgursky and Springer, 2007; Glazerman, Protik, Teh, Bruch, and Max, 2013). We use our estimates of relationships between short-run teacher effects and crime to ask how such policies affect future criminal outcomes, and explore whether policies that include non-cognitive measures could do better.

Concretely, we follow previous studies (Hanushek, 2011; Chetty et al., 2014b; Petek and Pope, 2016) that simulate how counterfactual teacher hiring policies can alter the students' future outcomes. These simulations replace teachers in the bottom vingtile of teacher quality with teachers of median quality, and see how such a swap will impact the distribution of student outcomes. We exploit the cognitive and non-cognitive posteriors from the multivariate model separately, and combinations of these two measures. The results are summarized by reporting the total change in student outcomes as a fraction of total crime.

[11] These effect estimates, notably the test score VA impacts, are different from those reported in Table 1.3 due to sample differences. Equation (1.14) is estimated on the sample of students with observed test score VA and future suspension VA. Equation (1.8) only restricted to students with the single observed teacher quality measure.

Table 1.6 summarizes the results of this simulation exercise. These estimates report the aggregate reduction in crime from the above teacher personnel policy a percentage of the total amount of each type of criminal outcome. We use our teacher effects from our multivariate random effects model in our simulation.

The first row shows how rates of first-time crime at age 20 are affected in each simulation. The results show very small reductions in crime when using test score evaluations of teachers - we would reduce all charges fall by approximately 0.019%, and all felony charges by 0.057%. We see similarly modest impacts of a test score hiring policy on the reductions of criminal convictions.

The second row performs the same experiment using future suspension effects, the measure of non-cognitive value-added with the largest effects on crime rates. Compared to a policy based on cognitive measures, this policy improves crime reduction by an order of magnitude. Aggregate charges fall by 0.27% and felony charges fall by 0.55%. Moreover, since cognitive value-added is conditionally uncorrelated with crime effects, policies that combine the two measures do no better than policies that use the non-cognitive measure alone. Moreover, since cognitive value-added is conditionally uncorrelated with crime effects, policies that combine the two measures do no better than policies that use the non-cognitive measure alone. they appear to be sufficiently capture any cognitive impacts of teachers' cognitive impacts on crime. The third row calculates the changes to aggregate crime using both future suspension and reading VA in this policy. The reductions in total crime using this combination measure are almost identical to the results of a policy using only future suspensions. The dominant feature of crime reduction comes from the non-cognitive impacts of teachers.

To provide a benchmark for the maximum possible impact of this teacher hiring policy, we also simulate a counterfactual hiring policy using "crime VA", in which we estimate (1.1) using crime at age 20 as the dependent variable. The results of this simulation have been reported in the final row. As explained in Section 1.4, these "crime VA" estimates will likely overestimate the causal effect of teachers on crime, and therefore will bound the extent to which teacher employment policies can be used to change crime. The estimates indicate maximum, a teacher hiring policy replacing the bottom 5% of teachers could reduce 0.26% of all criminal charges and 0.68% of felony charges. This means that policy simulations using non-cognitive VA are able to generate over half of all possible crime reductions through improved teacher quality.

Past studies have shown that test-score based personnel policies lead to substantial wage increases for students (Hanushek, 2011; Chetty et al., 2014b). The above simulations indicate that teacher hiring policies can also be used to reduce crime, although this can only be achieved if personnel decisions also incorporate non-cognitive teacher quality. Given the high social value of reduced crime (Heckman et al., 2010a), these estimates suggest substantial social welfare gains by using policies that combine multiple measures of teacher effectiveness (Neal, 2011).

1.8 Conclusion

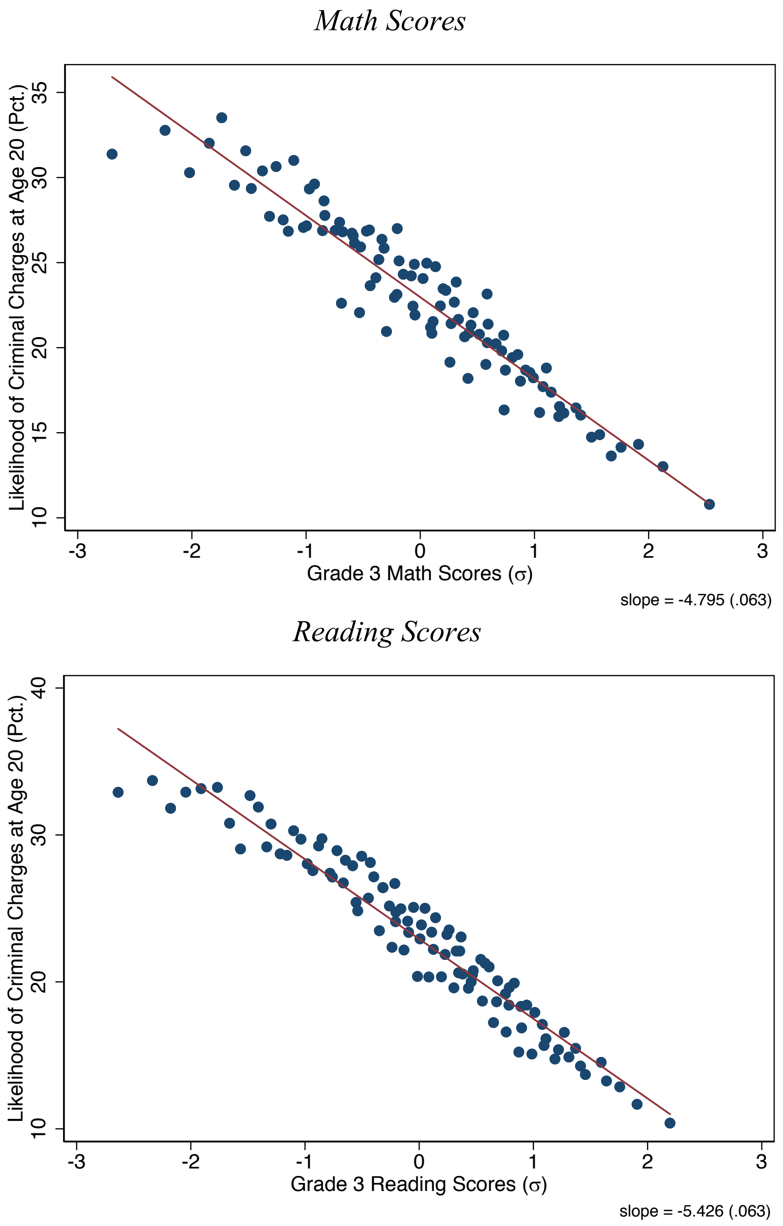
The literature on educational interventions and teacher quality in particular show that improved educational effectiveness has profound impacts on students. Our findings add to this constellation of findings, revealing that high-quality teachers also reduce crime. Moreover, we provide evidence that teachers mainly affect their students' crime propensity through a non-cognitive channel, and that these non-cognitive effects are unrelated to teachers' effects on test scores.

These results indicate that incentive-pay involving test-score measures alone will fail to maximize social returns for teachers, because such schemes will not incentivize any crime-reducing behavior among teachers. Moreover, test score performance incentives may crowd out teachers' development of their students' non-cognitive skills (Neal, 2011). While test score based incentives fall short of maximizing social welfare, future work is required to take into account the substantial heterogeneity in social costs of crime by offense type.

Pay performance schemes involving outcomes other than test scores are fairly new, and the optimal design and effectiveness of such merit pay schemes are open questions. The behavioral measures used to evaluate teachers in this paper, future suspensions and absences, are limited in their scope due to their indirect estimates of ability, and due to the change in interpretation depending on district and school policy. More direct assessments of changing non-cognitive ability of students, such as tests for executive function and effort, may recover more precise estimates of teacher impacts on non-cognitive outcomes (Moffitt, Arseneault, Belsky, Dickson, Hancox, Harrington, Houts, Poulton, Roberts, Ross, Sears, Thomson, and Caspi, 2011; Araujo, Carneiro, Cruz-Aguayo, and Schady, 2016). We leave this area open for future work.

1.9 Figures

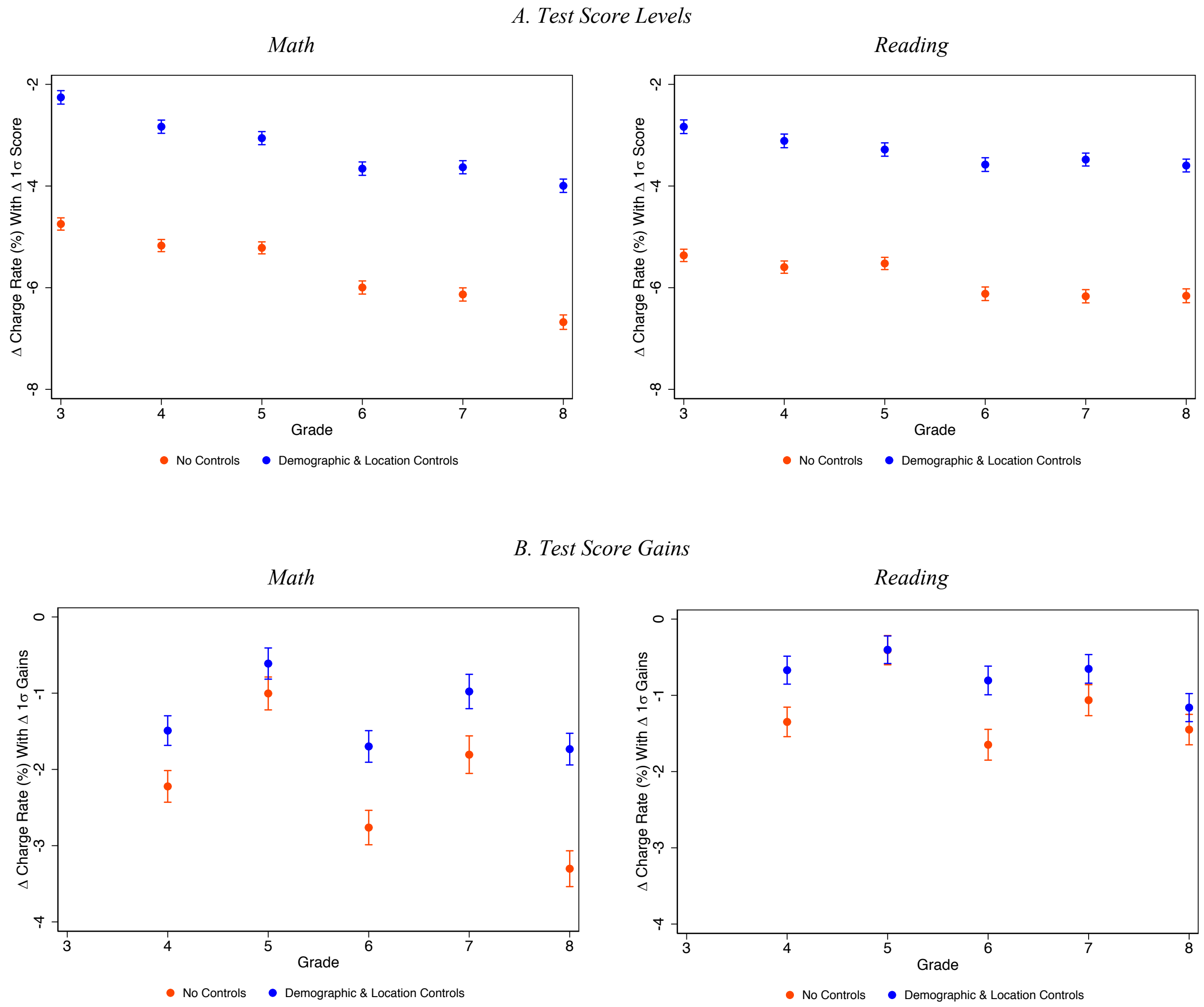
Figure 1.1: Relationships Between Grade 3 Test Scores and Criminal Charge Rates


Math Scores

Reading Scores

Notes: These graphs display binned scatter plots of criminal charge rates at age 20, in percentage points, as a function of third grade test scores. Test scores are standardized to have mean zero and standard deviation one by subject and year. These figures are generated by taking centiles of the test score distribution, calculating the average charge rate and test score average within each bin, and plotting these averages for each binned centile. Lines represent bivariate ordinary least squares regressions of a criminal charge indicator on test scores. The sample includes all third graders enrolled in schools with 25 or more students that were at least 20 years old in 2015.

Figure 1.2: Relationships Between Test Scores and Age 20 Crime Rates



A. Test Score Levels

Math

Reading

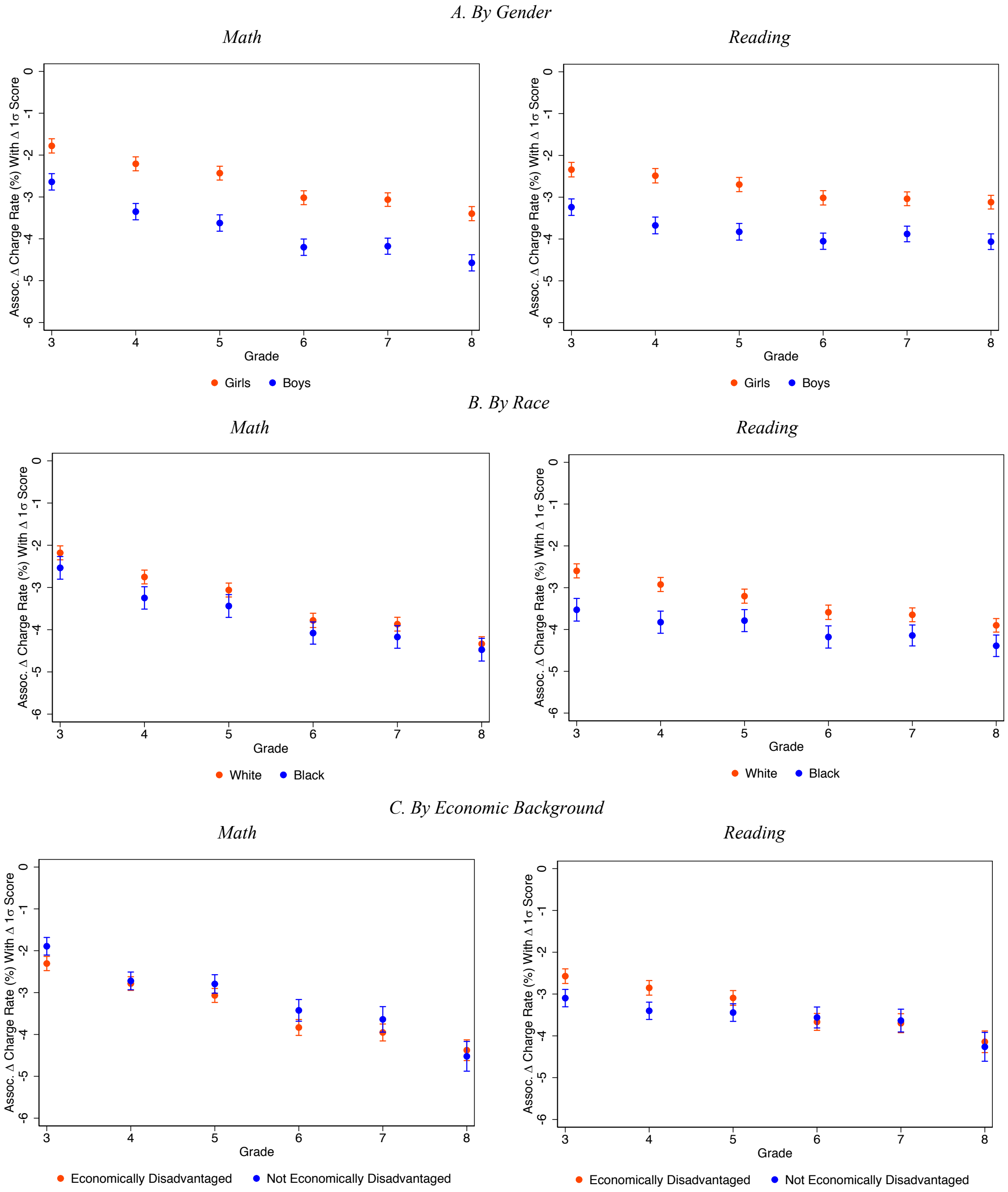
B. Test Score Gains

Math

Reading

Notes: This figure displays the association between the likelihood of criminal charges and test scores for grades 3 through 8. Panel A plots slope coefficients from regression of an indicator for criminal charges at age 20 on the test scores fit separately by grade. Panel B displays coefficients from regressions of a crime indicator on the change in test scores from the previous year. These coefficients are multiplied by 100. Each graph displays slopes from bivariate regressions (red points) and slopes from regressions with the following additional controls (blue points): gender, race, special education, limited english proficiency, economic disadvantage status, parental education, on-time grade progression from the previous year, contemporaneous suspensions and absences, and home tract and school fixed effects. Error bars display 95% confidence intervals, using standard errors that are clustered at the school-year level.

Figure 1.3: Relationships Between Test Scores and Age 20 Crime Rates By Demographics



A. By Gender

Math *Reading*

B. By Race

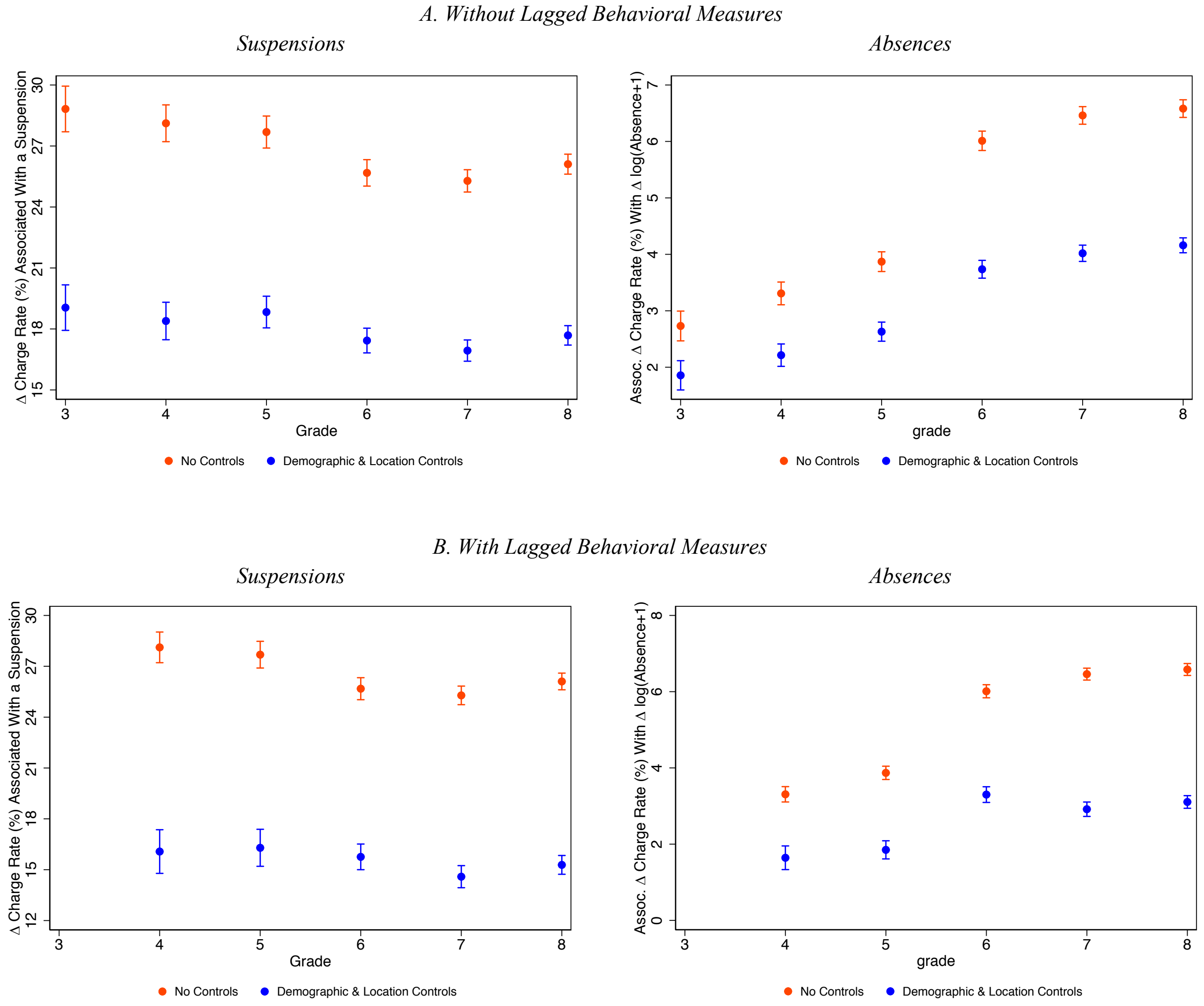
Math *Reading*

C. By Economic Background

Math *Reading*

Notes: This figure displays the associations between criminal charges and test scores for different subpopulations in grades 3 through 8. Points are slope coefficients from regressions of an indicator for a criminal charge by age 20 on test scores and control variables. The control variables are the same as the controls used in the regressions from Figure 2. Panel A plots the coefficients for boys and girls, Panel B for black students and white students, and Panel C for students who are economically disadvantaged and non-disadvantaged.

Figure 1.4: Relationships Between Behavioral Measures and Age 20 Crime Rates



A. Without Lagged Behavioral Measures

Suspensions

Absences

B. With Lagged Behavioral Measures

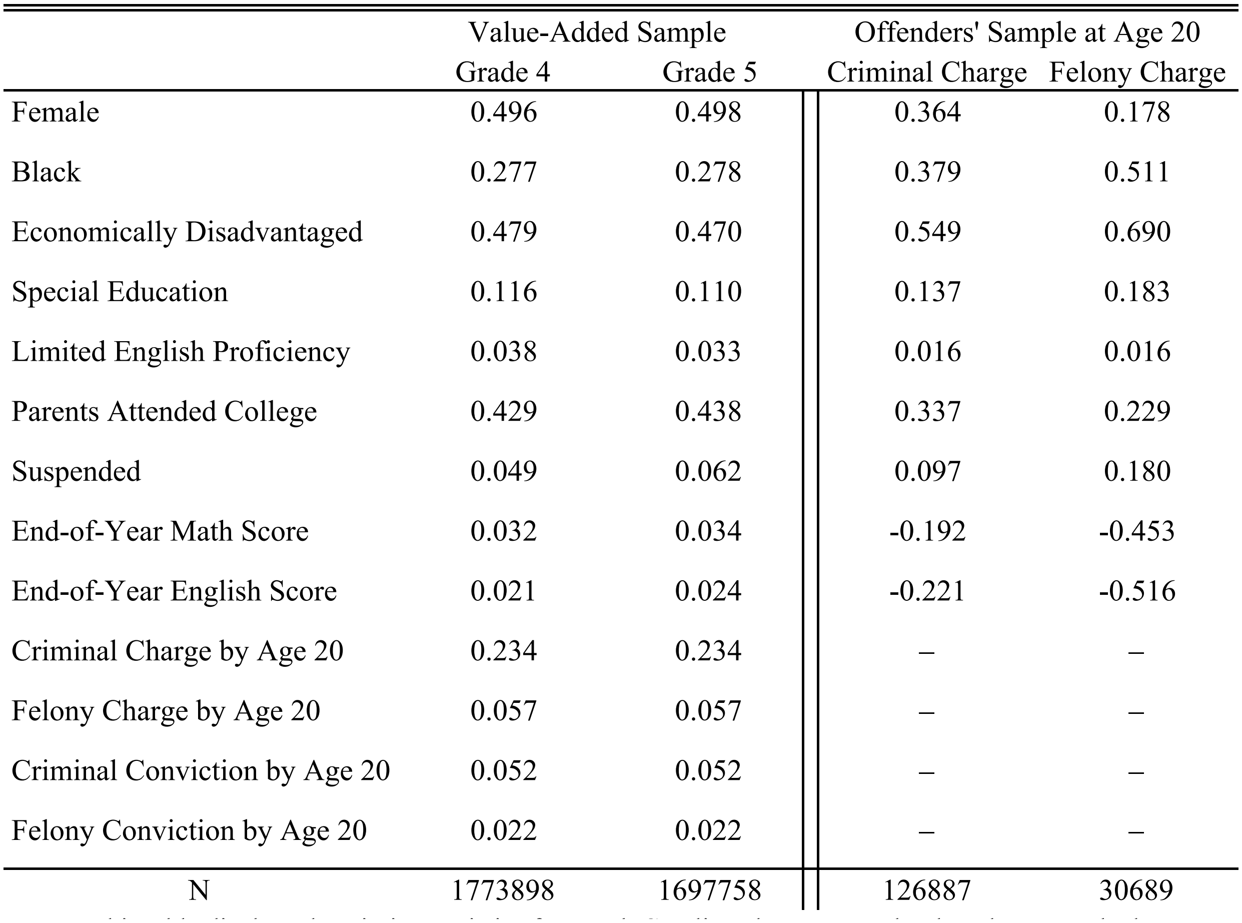
Suspensions

Absences

Notes: This graph displays associations between the likelihood of criminal charges and behavioral outcomes in grades 3 through 8. Left-hand panels show coefficients from regressions of a charge indicator on a suspension indicator, and right-hand panels show coefficients from regressions of a charge indicator on the log of one plus the number of absences. As in Figure 2, the graphs display both bivariate regression coefficients and coefficients from models that control for gender, race, special education, limited english proficiency, economic disadvantage status, parental education, on-time grade progression from the previous year, contemporaneous test scores, and home tract and school fixed effects. Panel B adds lagged behavioral measures as controls. Each regression is estimated separately by grade. Error bars display 95% confidence intervals.

1.10 Tables

Table 1.1: Summary Statistics



	Value-Added Sample		Offenders' Sample at Age 20	
	Grade 4	Grade 5	Criminal Charge	Felony Charge
Female	0.496	0.498	0.364	0.178
Black	0.277	0.278	0.379	0.511
Economically Disadvantaged	0.479	0.470	0.549	0.690
Special Education	0.116	0.110	0.137	0.183
Limited English Proficiency	0.038	0.033	0.016	0.016
Parents Attended College	0.429	0.438	0.337	0.229
Suspended	0.049	0.062	0.097	0.180
End-of-Year Math Score	0.032	0.034	-0.192	-0.453
End-of-Year English Score	0.021	0.024	-0.221	-0.516
Criminal Charge by Age 20	0.234	0.234	–	–
Felony Charge by Age 20	0.057	0.057	–	–
Criminal Conviction by Age 20	0.052	0.052	–	–
Felony Conviction by Age 20	0.022	0.022	–	–
N	1773898	1697758	126887	30689

Notes: This table displays descriptive statistics for North Carolina elementary school students matched to criminal records. The first two columns report mean demographics of fourth and fifth grade students, restricted to individuals with observed math and reading test scores in the relevant grade. Column (3) reports mean characteristics for our sample of individuals charged with a crime by age 20. Column (4) reports mean characteristics for individuals charged with a felony by age 20.

Table 1.2: Distribution of Teacher Value-Added

Skill Type	Outcome	Mean	Standard Deviation of Teacher Effects	
			(1)	(2)
Cognitive	Math Scores	0.055	0.397	0.171
			(0.002)	(0.001)
	Reading Scores	0.040	0.360	0.095
			(0.002)	(0.001)
Non-cognitive	Contemporaneous Suspensions	0.056	0.053	0.031
			(0.001)	(0.000)
	Future Suspensions	0.108	0.071	0.034
			(0.001)	(0.001)
	Contemporaneous log(Absences+1)	1.670	0.156	0.083
			(0.001)	(0.001)
	Future log(Absences+1)	1.722	0.158	0.108
			(0.002)	(0.001)
Criminal Charges by Age 20	Any Charge	0.233	0.048	0.031
			(0.001)	(0.001)
	Felony Charge	0.056	0.024	0.011
			(0.000)	(0.000)
Controls				X

Notes: This table displays the standard deviations of teacher value-added on outcomes for fourth and fifth-grade students. The row indexes which value-added measure we are estimating. Standard deviations are maximum likelihood estimates of σ_α from Equation (**??**). Column (1) estimates this model without any controls. Column (2) adds controls for cubics in lagged math and reading test scores, sex, race, special education, limited english proficiency, economic disadvantage status, parental education, on-time grade progression from the previous year, lagged suspensions and absences, school-level averages of each of these variables, and grade and year dummies. Samples are restricted to teachers observed in multiple years with at least 25 students. Samples for cognitive and non-cognitive outcomes are also restricted to students with a lagged measure of the outcome variable. Future non-cognitive outcomes are measured in sixth grade. Mean values of the outcomes are also reported.

Table 1.3: Impacts of Value-Added on Crime

	Charges at Age 20		Convictions at Age 20	
VA Measure	Any	Felony	Any	Felony
	(1)	(2)	(3)	(4)
Math Scores	-0.028	-0.019	-0.084	-0.039
	(0.049)	(0.027)	(0.026)	(0.017)
Reading Scores	-0.104	-0.033	-0.087	-0.032
	(0.050)	(0.027)	(0.026)	(0.017)
Contemporaneous Suspensions	0.191	0.026	0.188	0.026
	(0.072)	(0.039)	(0.037)	(0.025)
Future Suspensions	0.417	0.243	0.247	0.130
	(0.079)	(0.043)	(0.041)	(0.027)
Contemporaneous log(Absences+1)	-0.003	0.076	0.052	0.037
	(0.105)	(0.057)	(0.053)	(0.035)
Future log(Absences+1)	0.480	0.265	0.149	0.153
	(0.102)	(0.055)	(0.051)	(0.034)

Notes: This table reports estimates of relationships between crime outcomes and teacher cognitive and non-cognitive value-added. Estimates come from regressions of a crime indicator on posterior mean predictions of teacher value-added (Equation (1.8)). Coefficients are multiplied by 100. Value-added measures are standardized so that the reported estimates can be interpreted as the impact of a one-standard deviation change in value-added on the likelihood of committing a criminal offense by age 20, in percentage points. Column (1) reports the teacher impacts on any criminal charges, Column (2) reports the impacts on felony charges, Column (3) on any convictions, and Column (4) on felony convictions. Standard errors are reported in parentheses and are clustered at the teacher level.

Table 1.4: Multivariate Distributions of Value-Added

	Univariate Model		Multivariate Model	
	Reading Scores	Future Suspensions	Reading Scores	Future Suspensions
	(1)	(2)	(3)	(4)
Mean Outcome	0.040	0.108	0.098	0.103
Impact of a Std. Dev. Increase in α on Outcome	0.095	0.034	0.097	0.060
	(0.001)	(0.001)	(0.001)	(0.001)
Correlation in Teacher Effects on Cogntive and Non-Cognitive Skills (ρ_{α})	-		0.027	
			(0.085)	
Correlation in Student Unobservables (ρ_{ϵ})	-		-0.063	
			(0.004)	

Notes: These numbers reflect and compare the parameter estimates of teacher effects from the univariate model outlined in Section 1.4 and the multivariate random effects model in Section 1.6. The second row displays the impact of a standard deviation increase in the random effect on the given VA measure. For columns (1), (2), and (3), this value corresponds to the standard deviation of the teacher value-added (VA) for the given outcome, which is our estimates of σ_α. In Column (4), as α is a latent effect in the multivariate model, this corresponds to the change in future suspension probability due to a one standard deviation change in standard deviation in at mean characteristics. The third row displays the correlations between reading and future suspension VA, or ρ_α in the multivariate model. The fourth row displays our estimate of ρ_ϵ, the correlation coefficient between student unobservables for reading and future suspensions.

Table 1.5: Impacts of Total Elementary Value-Added on Criminal Activity at Age 20

Criminal Outcome		VA Measure	Multivariate Model			Univariate Model		
			(1)	(2)	(3)	(4)	(5)	(6)
Charges	Any	Reading Scores	0.055		0.053	0.117		0.116
			(0.114)		(0.114)	(0.123)		(0.123)
		Future Suspensions		0.601	0.601		0.463	0.463
				(0.110)	(0.110)		(0.135)	(0.135)
	Felony	Reading Scores	0.039		0.038	0.103		0.103
			(0.058)		(0.059)	(0.062)		(0.062)
		Future Suspensions		0.297	0.297		0.275	0.275
				(0.058)	(0.058)		(0.075)	(0.075)
Convictions	Any	Reading Scores	-0.054		-0.055	-0.064		-0.068
			(0.061)		(0.060)	(0.065)		(0.065)
		Future Suspensions		0.399	0.399		0.348	0.348
				(0.062)	(0.062)		(0.077)	(0.077)
	Felony	Reading Scores	-0.016		-0.016	-0.016		-0.014
			(0.035)		(0.035)	(0.039)		(0.039)
		Future Suspensions		0.159	0.159		0.133	0.133
				(0.035)	(0.035)		(0.044)	(0.044)

Notes: This table reports our estimates of equation (1.14). Estimates are reported using the sample used to estimate the correlated random effects model, which restricts our sample to teachers who have taught at least 25 students with observations of of both outcomes paired with their respective lags. Column (1) through (3) display estimates using our VA estimates from our correlated random effects model, our preferred specification. Columns (1) and (2) only include cognitive and non-cognitive VA measures, respectively, and Column (3) includes both measures. Column (4) through (6) uses the univariate VA estimates in place of our multivariate estimates. Similarly, Columns (4) and (5) only include cognitive and non-cognitive VA measures, respectively, and Column (6) includes both measures. We test this model using four measures of crime: whether or not your were charged or convicted of any crime or any felony crime by age 20. Standard errors are clustered at the teacher level.

Table 1.6: Policy Simulations

Value-Added Measure	Charges		Convictions	
	All	Felonies	All	Felonies
	(1)	(2)	(3)	(4)
Reading Scores	-0.017	-0.050	-0.038	-0.066
Future Suspensions	-0.268	-0.552	-0.436	-0.798
Reading Scores and Future Suspensions	-0.269	-0.552	-0.458	-0.802
Crime (theoretical limit)	-0.382	-0.676	-1.399	-0.925

Notes: This table reports predicted effects of policies that replace the bottom 5% of teachers according to some measure of teacher value-added with an average teacher. Estimates are predicted changes in the likelihood of a crime by age 20, in percentage points multiplied by 100. Value-added estimates come from the multivariate random effects model in Table 5. The reading score measure uses only test score value-added, while the future suspension measure uses only non-cognitive value-added. The "reading score and future suspension" measure combines cognitive and non-cognitive value-added to predict a teacher's effect on crime, and replaces the bottom 5% according to this metric. The crime value-added simulation uses a direct measure of a teacher's effect on future crime. This policy is infeasible in practice because crime outcomes are only observed many years into the future.

1.11 Appendix: Data Construction

Education Data

Our sample was constructed as described in Section 1.2. However, our panel of outcomes is not continuous, and both our test score and behavioral outcomes ar emissing for several years of our panel from 1995 to 2011. The test score data is mostly complete, although in the 1995-96 cohort, we do not observe grade 5 test scores, meaning that we can only evaluate test score quality for fourth grade teachers in these years. The attendance data begins in the 2003-04 school year, which is the first year we can evaluate teachers along this metric.

The discipline records exist beginning in the 2000-01 school year, with the exception of the 2004-05 school year, in which there are no student IDs to match the suspension records to the other education records. However, these records were more difficult to process than our other outcomes. We focused on out-of-school suspensions, as these by law are forced to be reported (detentions and in-school suspensions are not), and are relatively frequent in nature (approximately 10% of sixth graders are suspended at least once in a given year). However, this legal enforcement did not ensure that the records were kept in the administrative data. Before the 2007-08 school year, approximately 50% per year, did not have any students appear in the disciplinary data, whereas by 2007-08, over 98% of schools had at least one student appear in the discipline records each year. While it is possible that these records are accurate and there was simply a major uptick in suspensions in the mid-2000s, we believe this indicates that schools varied widely in their reporting of disciplinary records until the 2007-08 school year. For this reason, we only estimate suspension VA for teachers attending schools that suspended at least one student in the given year. We therefore interpret our suspension effects as an intensive margin effect, meaning that our effects estimate the variation in your likelihood of being suspend, given that your school is willing to suspend its students.

Matching Students to Teachers

In our main specification, we match elementary school students to teachers based on the test score proctor. This match is only successful in identifying younger grades, as students typically only have one teacher.

For middle school teachers, we utilize the course membership files, which exist from 2007 through 2016 for all grades. These records identify teacher-student pairs for all unique course codes, allowing us to identify subject-specific cognitive ability gain for math and reading classes for teachers in grades 4 though 8 for 2007-2016. This sample has the advantage of being more exact – we know *exactly* who each instructor for each course is, whereas in our main analysis, we only observe the test score proctor, forcing us to err on the conservative side and omit several classrooms that had proctors that we could not certify were teachers. However, these direct classroom match limits the duration of our long-run crime outcomes that we can observe for these students, which is why we focus on the former.

Table A1 summarizes the sample of middle school students matched using these course membership files. We also made (omitted from the text) a similar table for elementary school student to compare observables. These descriptives are very similar, with the exception of the limited English proficiency (LEP) rate. We see almost double the population of LEP students in our course membership sample as we do in the test proctor sample. This difference appears to come from two key differences: the share of LEP students in North Carolina public schools has increased substantially in recent years, and limiting to test takers in the test proctor sample causes the LEP fraction to drop by about one-third, suggesting that LEP students were not taking these exams in earlier years of the sample.

Education and Crime Data Merge

Our criminal records were collected from two sources: The North Carolina District and Superior Courts (DSC) and the Department of Public Safety (DPS). The former records included any time there was a criminal charge filed. These records begin in 1996 and include all severe misdemeanor and felony cases; beginning in 2005 and continuing through 2015, these records then also included *all* misdemeanor and felony charge cases. The latter records span back through 1970 and include all instances when you were sentenced for a criminal conviction that led to some sort of mandated supervision, whether it be parole or incarceration, an outcome that was required by law for all felonies and severe misdemeanors. We focus on the years 2005-2015 in our crime analysis, as these are the richest crime records that we have, although the North Carolina Education Research Data Center (NCERDC), who manages the administrative education records for the state, attempted to merge *all* of the unique criminal records, totaling 5,495,303 records, to the education data.

The NCERDC performed the match using three characteristics of the offender: name, date of birth (DOB), and last four digits of the social security number (SSN4). Approximately 60% of our merged arrest records happened on using all three criteria, about 35% of the remaining records were merged using name matches and DOB alone, with the remaining fraction of the records either due to exact names and SSN4, or using approximate names, DOB, and SSN4 together. When we say approximate names, we included exact matches, matches using common nicknames, manual misspellings, and SPEDIS scores of distance between the names in the two source; we considered matches where the SPEDIS score if both first name and last names had scores no greater than 40, or if either name had a score no greater than 35.

Given this procedure, we were able to match 1,170,683 of the approximate 5.5 million records to education records. The vast majority (approximately 75%) of the criminal records that were unmatched were of individuals born before 1978, meaning that most of these offenders would have graduated or left high school before appearing in our sample of educational records, which begin in 1994-95 school year.

To verify the quality of our match, we used state-level aggregates for our cohorts to verify our match numbers. According to North Carolina's Department of Vital Statistics, 104439 individuals were born in the state. In our records, we find 10683 of people born in 1990 had a felony charge by age 25. We were able to match 10259 of these individuals to first-time third graders in the state of North Carolina, at a rate of 10.2%. Overall, we find that in our sample, the felony charge rate by age 25 is 11.4%, slightly higher than the average. There are two possible explanations for this. First, our sample only looks at the outcomes of public school students. While this accounts for almost 90% of the total population of students in North Carolina, our sample is more at risk and thus may have higher rates of offending. It is also possible that part of this could be due to migration. North Carolina has one of the highest positive net migration rates of any state in the country; if young children move into the state and commit crimes at a higher rate than native born individuals, we will see an uptick in numbers.

We are not overly concerned with either of these, as the threats to external validity given the coverage of our data our low, and net inflow of migrants will not threaten our estimates. The main concern would be migration *out* of the state that happens systematically in a way related to teacher assignment and criminality; we are labeling individuals who do not appear in the criminal justice records as non-offenders, but if individuals who left the state committed crimes in other locations, our results would be biased. However, given the extremely accurate matching of our 1990 felony charge counts to aggregate samples, we do not believe this will be a problem.

Moreover, our criminal offending rates are quite close to numbers are similar to numbers found in other samples. In particular, Brame et al. (2012) show an arrest rate of 24.0% nationally by age 20 when the National Longitudinal Survey of Youth in 1997, a nationally representative sample of adolescents. Moreover, in unpublished work, Rose and Shem-Tov (2018) find similar rates of offending in Washington State.[12]

1.12 Appendix: Relationship Between 2SLS and Value-Added

Baseline Model

In this section, we formally show the relationship between 2SLS estimators value-added estimates. This will be a simplified version of the VA framework, but can readily be extended to the model in Section 1.4. Ultimately, the goal is to determine the effect of a teacher's VA $\alpha_{j(i)}$ on outcome Y_i (e.g., earnings at age 28, or criminal charges by age 20). First, a fixed effect model is used to recover estimates of teacher impacts on achievement measure A_i, such as test scores, after controlling for other characteristics of the student X_i, as follows:

[12] Important to note: Billings et al. (2013) have a similar match pattern for a subset of our years within Charlotte-Mecklenburg county, although their estimated arrest rates are much lower than ours. This discrepancy is due to their sample construction, where they take a fixed panel of arrest dates and match multiple years of educational records to these arrests. We only consider charge rates by a given age, and restrict our sample to individuals whose entire criminal history until that age would be observed in the data.

$$A_i = X_i'\beta_0 + \alpha_{j(i)} + e_i \tag{1.15}$$

Ultimately, the goal is to estimate the following:

$$Y_i = X_i'\eta + \gamma\alpha_{j(i)} + u_i \tag{1.16}$$

Here, γ is the causal parameter of interest. $\alpha_{j(i)}$ is not observed, and is proxied with our estimate $\hat{\alpha}_{j(i)}$ from (1.15). However, due to concerns of measurement error attenuating the estimated teacher effects, we will use a shrunken estimate $f_j(\hat{\alpha})$ in place of α_j when estimating Equation (1.16), shown below:

$$Y_i = X_i'\eta + \gamma f_j(\hat{\alpha}_{j(i)}) + u_i \tag{1.17}$$

The shrinkage function $f_j(\cdot)$ depends on the type of shrinkage used.

Now consider a causal model described by:

$$Y_i = X_i'\beta_2 + \tilde{\gamma}A_i + \eta_i \tag{1.18}$$

Suppose we try to estimate $\tilde{\gamma}$ using teacher indicators as instruments. The first stage would be the regression depicted in (1.15). Substituting in, we get the following expression for the expression for Y:

$$\begin{aligned}
Y_i &= X_i'\tilde{\beta}_1 + \tilde{\gamma}\hat{A}_i + u_i && (1.19)\\
&= X_i'\tilde{\beta}_1 + \tilde{\gamma}(X_i'\hat{\beta}_0 + \hat{\alpha}_{j(i)}) + u_i && (1.20)\\
&= X_i'(\tilde{\beta}_1 + \tilde{\gamma}'\hat{\beta}_0) + \tilde{\gamma}\hat{\alpha}_{j(i)} + u_i && (1.21)\\
&= X_i'\beta^* + \tilde{\gamma}\hat{\alpha}_{j(i)} + u_i^* && (1.22)
\end{aligned}$$

But this is exactly the same model as in our VAM estimates whenever $f_j(\hat{\alpha}_{j(i)}) = \hat{\alpha}_{j(i)}$, or the model without any shrinkage. In other words, the causal parameter we are interested in is also identified by a 2SLS regression of the long-run outcomes on test score gains using teacher assignment as an instrument.

Shrinkage with constant n per teacher

A common shrinkage strategy is to use homoskedastic empirical Bayes assuming student errors and teacher quality are normally distributed, as shown below:

$$\begin{aligned}
e_i|\alpha_{j(i)} &\sim N(0, \sigma_e^2) && (1.23)\\
\alpha_j &\sim N(0, \sigma_\alpha^2) && (1.24)
\end{aligned}$$

This approach is somewhat difficult to implement in the presence of covariates. In the past, some researchers have residualized A_i on covariates, and then applied a shrinkage using these residualized measures (Chetty et al., 2014a). Doing so fails to account for the sorting of students to teachers based on observable characteristics, as described in Section 1.4.

We instead model $\alpha_{j(i)} \sim N(\bar{X}_j'\gamma, \sigma_\alpha^2)$, meaning that teacher value added can depend on average student characteristics. This allows us to avoid shrinking towards a "grand mean" of zero and shrink towards average covariate-dependent means instead.

First, we can estimate $\alpha_{j(i)}$ and γ using the following regression:

$$A_i = X_i'\beta_0 + \bar{X}_j'\gamma + e_i \tag{1.25}$$

We can then form estimates of $\tilde{\alpha}_{j(i)} = \alpha_{j(i)} - \bar{X}_j'\gamma$ as follows:

$$\hat{\tilde{\alpha}}_{j(i)} = \bar{\tilde{A}}_j \tag{1.26}$$

$$\bar{\tilde{A}}_j = \frac{1}{N}\sum_{i\in j} A_i - X_i'\hat{\beta}_0 - \bar{X}_j'\hat{\gamma} \tag{1.27}$$

The resulting shrunken value-added estimate for teacher j is:

$$V\hat{A}_j = (1-\lambda)\bar{X}_j'\hat{\gamma} + \lambda(\hat{\tilde{\alpha}}_{j(i)} + \bar{X}_j'\hat{\gamma}) \tag{1.28}$$

$$\lambda = \frac{\sigma_\alpha}{\sigma_\alpha + \sigma_e/N} \tag{1.29}$$

Suppose that instead of doing all this, we use $\bar{\tilde{A}}_j$ and $\bar{X}_j'$ as instruments in a 2SLS set-up. If we want the first stage to include some kind of shrinkage, we can use a version of the student-level leave-out mean $\bar{\tilde{A}}_{j(-i)} + \bar{X}_j'\hat{\gamma}$ along with $\bar{X}_j$ itself as instruments. The first stage regression is then:

$$A_i = X_i'\beta_0 + \bar{X}_j'\beta_1 + \delta(\bar{\tilde{A}}_{j(-i)} + \bar{X}_j'\hat{\gamma}) + e_i \tag{1.30}$$

$$= X_i'\beta_0 + \bar{X}_j'(\beta_1 + \delta\hat{\gamma}) + \delta\hat{\tilde{\alpha}}_{j(-i)} + e_i \tag{1.31}$$

$$= X_i'\beta_0 + \bar{X}_j'\hat{\gamma} + \delta\hat{\tilde{\alpha}}_{j(-i)} + e_i \tag{1.32}$$

where the last line follows from the fact that the leave out mean $\hat{\tilde{\alpha}}_{j(i)}$ is orthogonal to $\bar{X}_j$, so that regression is the same as one that omits the leave out mean all together. This implies that $\beta_1 + \delta\hat{\gamma} = \hat{\gamma}$ so that $\beta_1 = \hat{\gamma}(1-\delta)$.

After applying the Frisch-Waugh-Lovell Theorem to X_i and $\bar{X}_j$, δ will be given by the following expression:

$$\delta = \frac{cov(\tilde{\alpha}_j, \tilde{\alpha}_j + \frac{1}{N-1}\sum_{k\in j, k\neq i} e_k)}{var(\tilde{\alpha}_j + \frac{1}{N-1}\sum_{k\in j, k\neq i} e_k)} \tag{1.33}$$

$$= \frac{\sigma_\alpha}{\sigma_\alpha + \sigma_e/(N-1)} \tag{1.34}$$

$$= \tilde{\lambda} \tag{1.35}$$

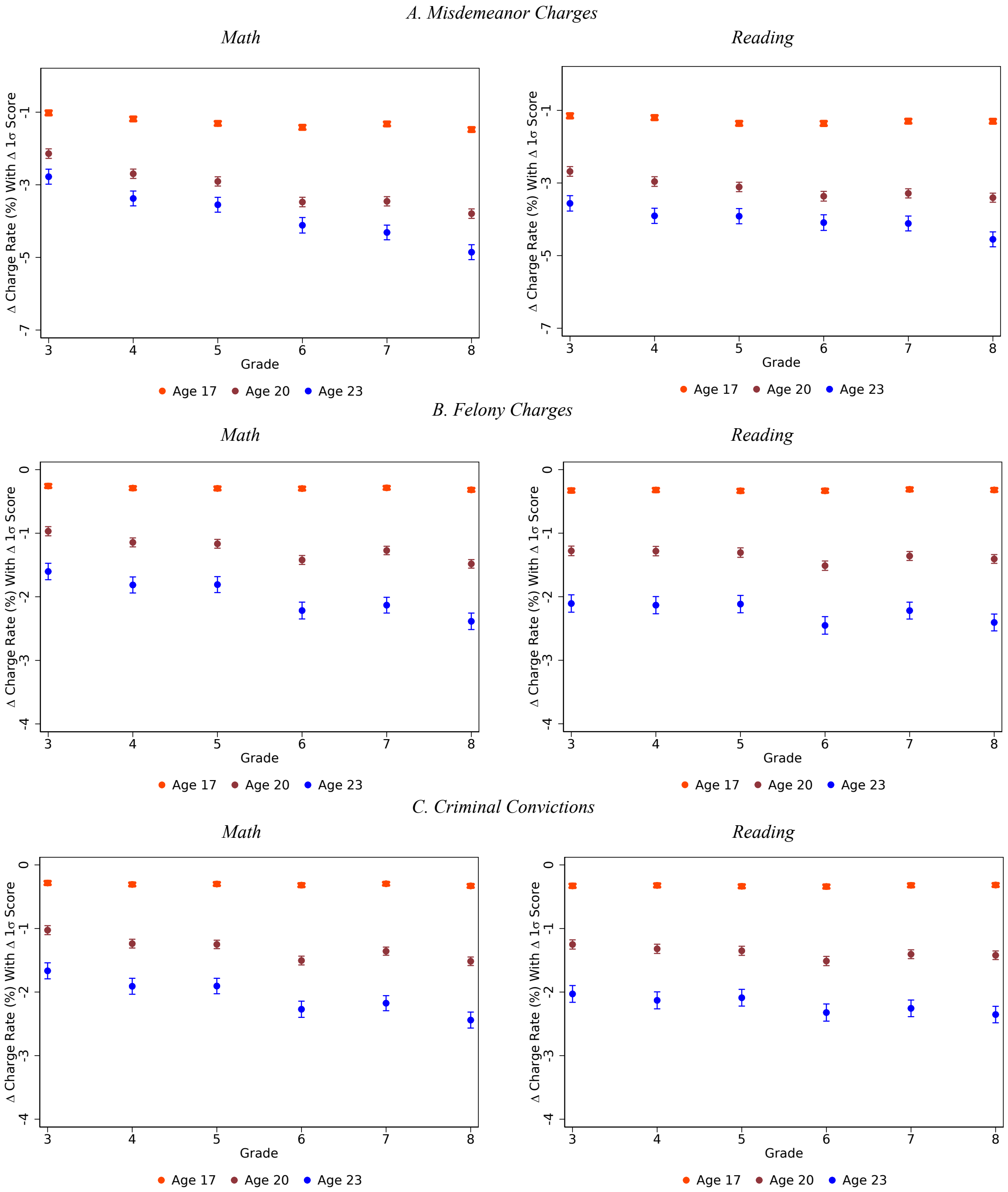
Thus, we can express Y as follows:

$$\begin{aligned}
Y_i &= X_i'\tilde{\beta}_1 + \tilde{\gamma}\hat{A}_i + u_i && (1.36)\\
&= X_i'\tilde{\beta}_1 + \tilde{\gamma}\left(X_i'\hat{\beta}_0 + \bar{X}_j'\beta_1 + \tilde{\lambda}(\bar{\bar{A}}_{j(-i)} + \bar{X}_j'\hat{\gamma})\right) + u_i && (1.37)\\
&= X_i'(\tilde{\beta}_1 + \tilde{\gamma}\hat{\beta}_0) + \tilde{\gamma}\left((1-\tilde{\lambda})\bar{X}_j'\hat{\gamma} + \tilde{\lambda}(\bar{\bar{A}}_{j(-i)} + \bar{X}_j'\hat{\gamma})\right) + u_i && (1.38)\\
&= X_i'(\tilde{\beta}_1 + \tilde{\gamma}\hat{\beta}_0) + \tilde{\gamma}\hat{VA}_j + u_i && (1.39)
\end{aligned}$$

The very last step is an approximation, since $\tilde{\lambda} \neq \lambda$, but should be very similar with a large sample. Note that this "jackknife" procedure is not quite the same as the one presented in Section 1.4. Namely, in order to avoid serial correlation of the classroom effect, we omit *cohorts* in our teacher estimates, not simply the individual in question. However, in parallel, we can expect the noise term in the corresponding $\tilde{\lambda}$ to be slightly larger. However, assuming the number of students is significantly larger than the number of students per classroom, which will be true in cases with many years of data.

1.13 Appendix: Additional Figures

Figure 1.5: Relationships Between Test Scores and Crime Rates By Age and Crime Severity

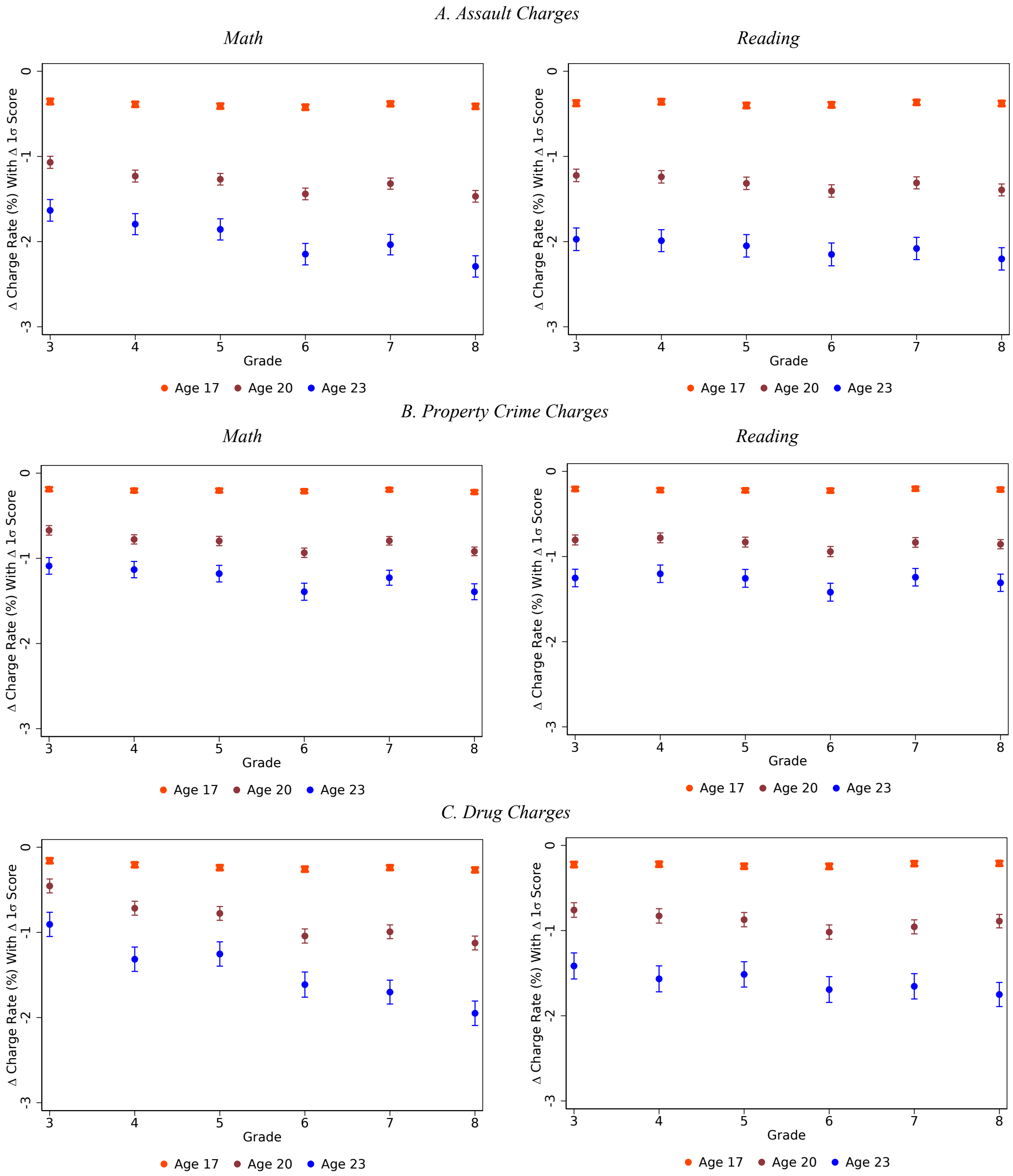

A. Misdemeanor Charges

B. Felony Charges

C. Criminal Convictions

Notes: This figure displays the associations between criminal charges and test scores for different subpopulations in grades 3 through 8. The figures vary the criminal charge of interest, and the age of the first offense. Panel A plots the coefficients for misdemeanor charges, Panel B for felony charges, and Panel C for criminal convictions. All images plot the associations for first criminal charges of the given type by ages 17, 20, and 23. Points are slope coefficients from regressions of an indicator for a criminal charge by a given age on test scores and control variables. The control variables are the same as the controls used in the regressions from Figure 2. Each regression is estimated separately by grade. Error bars display 95% confidence intervals.

Figure 1.6: Relationships Between Test Scores and Crime Rates by Age and Crime Type



A. Assault Charges

Math *Reading*

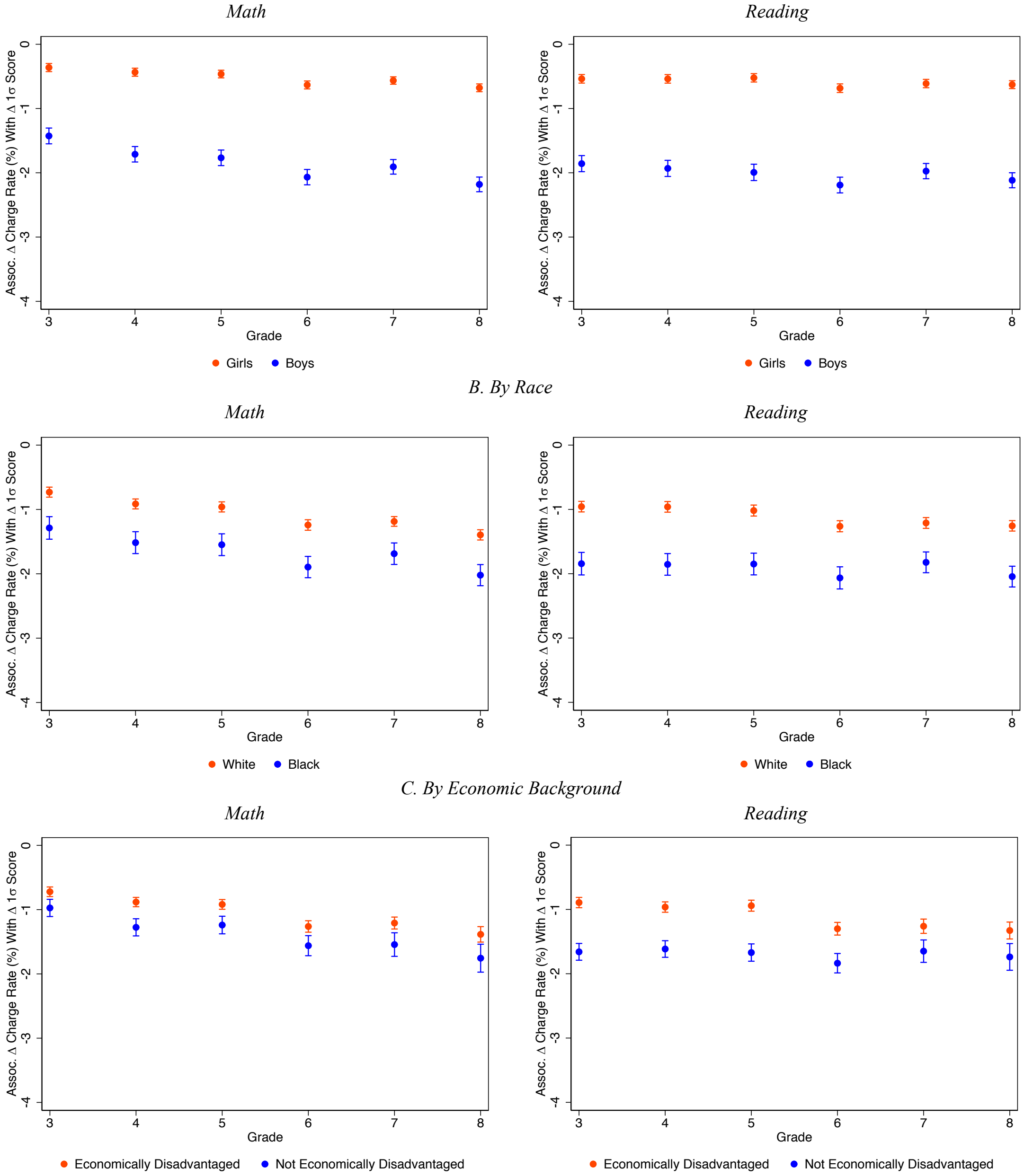
B. Property Crime Charges

Math *Reading*

C. Drug Charges

Notes: This figure displays the associations between criminal charges and test scores in grades 3 through 8. The figures vary the criminal charge of interest and the age of the first offense. Panel A plots the coefficients for assault charges, Panel B for property crime charges, and Panel C for drug charges. All images plot the associations for first criminal charges of the given type by ages 17, 20, and 23. Points are slope coefficients from regressions of an indicator for a criminal charge by a given age on test scores and control variables. The control variables are the same as the controls used in the regressions from Figure 2. Each regression is estimated separately by grade. Error bars display 95% confidence intervals.

Figure 1.7: Relationships Between Test Scores and Age 20 Felony Charge Rates Across Demographic Groups



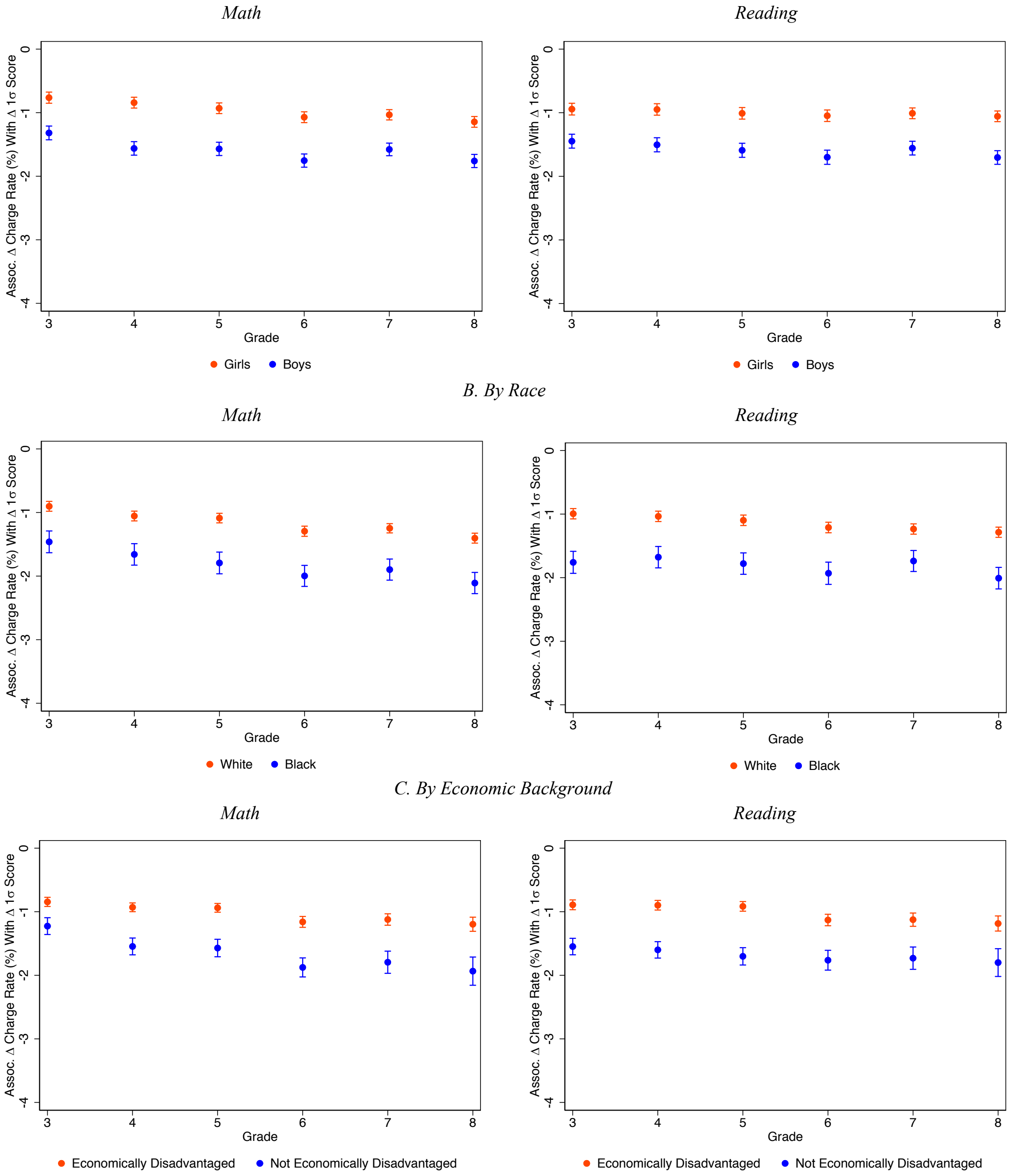
A. By Gender

B. By Race

C. By Economic Background

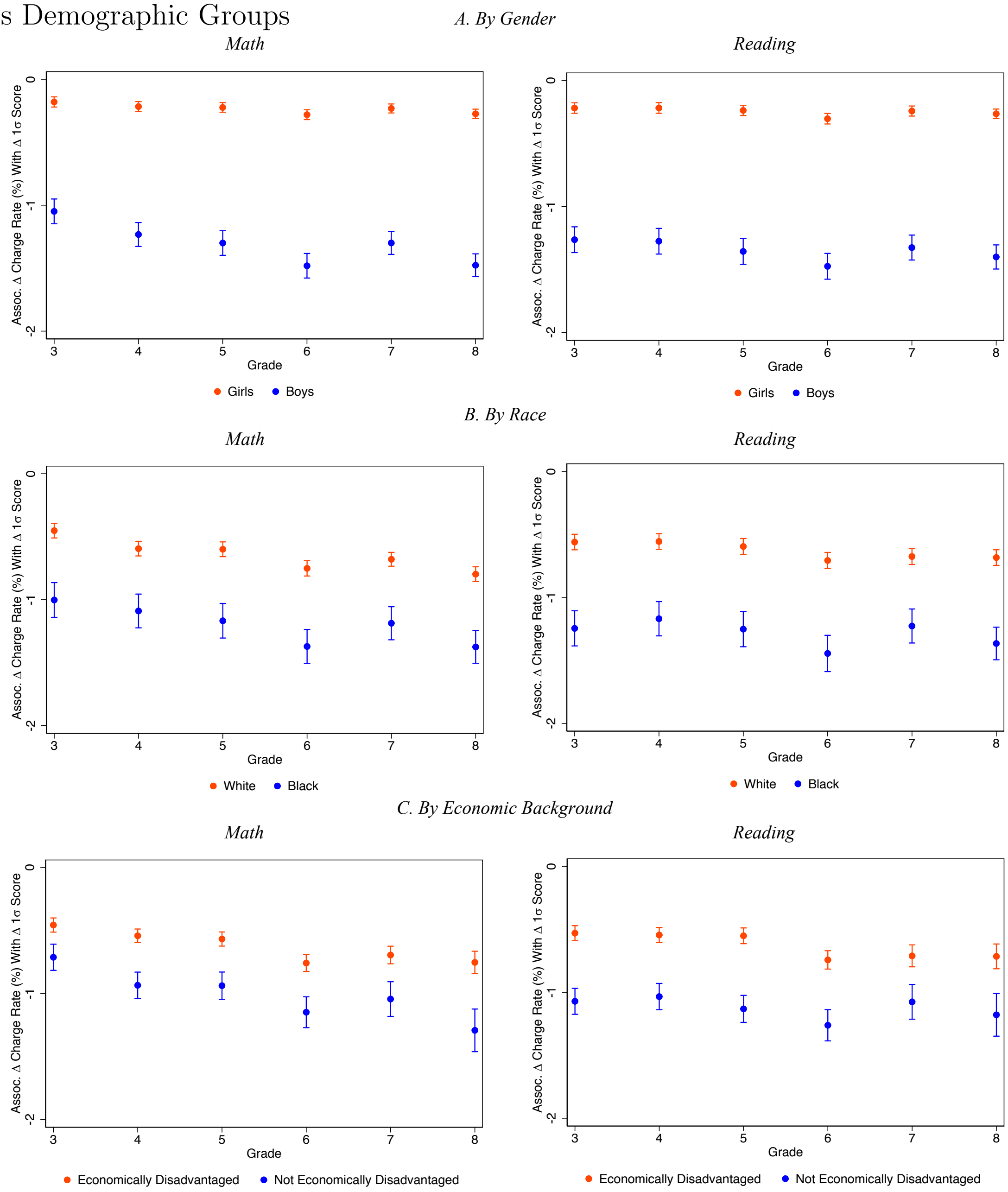
Notes: This figure displays the associations between felony criminal charges and test scores for different subpopulations in grades 3 through 8. Points are slope coefficients from regressions of an indicator for a felony criminal charge by age 20 on test scores and control variables. The control variables are the same as the controls used in the regressions from Figure 2. Panel A plots the coefficients for boys and girls, Panel B for black students and white students, and Panel C for students who are economically disadvantaged and non-disadvantaged. Each regression is estimated separately by grade. Error bars display 95% confidence intervals.

Figure 1.8: Relationships Between Test Scores and Age 20 Assault Charge Rates Across Demographic Groups



Notes: This figure displays the associations between assault charges and test scores for different subpopulations in grades 3 through 8. Points are slope coefficients from regressions of an indicator for an assault charge by age 20 on test scores and control variables. The control variables are the same as the controls used in the regressions from Figure 2. Panel A plots the coefficients for boys and girls, Panel B for black students and white students, and Panel C for students who are economically disadvantaged and non-disadvantaged. Each regression is estimated separately by grade. Error bars display 95% confidence intervals.

Figure 1.9: Relationships Between Test Scores and Age 20 Property Crime Charge Rates Across Demographic Groups



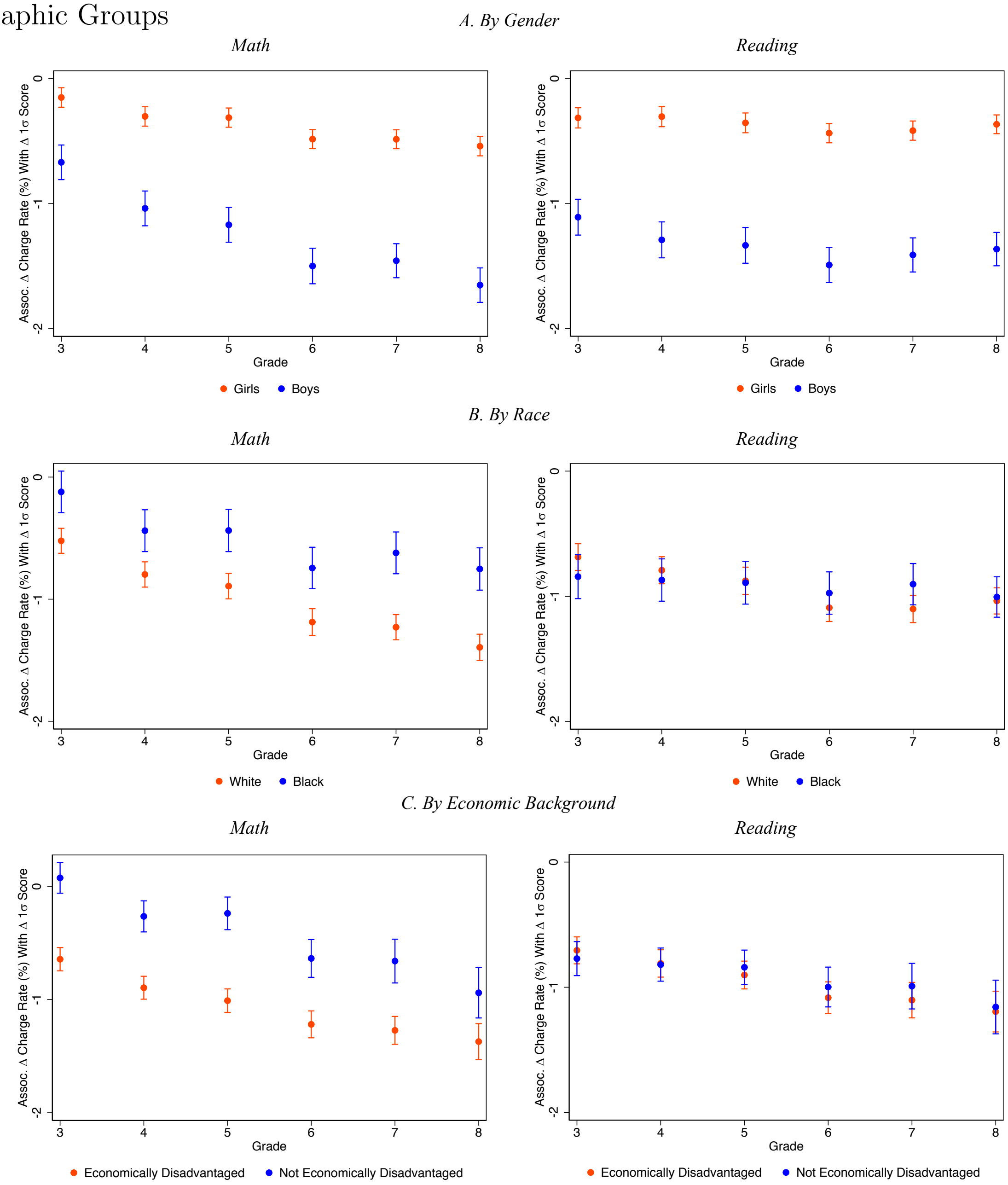
A. By Gender

B. By Race

C. By Economic Background

Notes: This figure displays the associations between property crime charges and test scores for different subpopulations in grades 3 through 8. Points are slope coefficients from regressions of an indicator for a property crime charge by age 20 on test scores and control variables. The control variables are the same as the controls used in the regressions from Figure 2. Panel A plots the coefficients for boys and girls, Panel B for black students and white students, and Panel C for students who are economically disadvantaged and non-disadvantaged. Each regression is estimated separately by grade. Error bars display 95% confidence intervals.

Figure 1.10: Relationships Between Test Scores and Age 20 Drug Crime Rates Across Demographic Groups

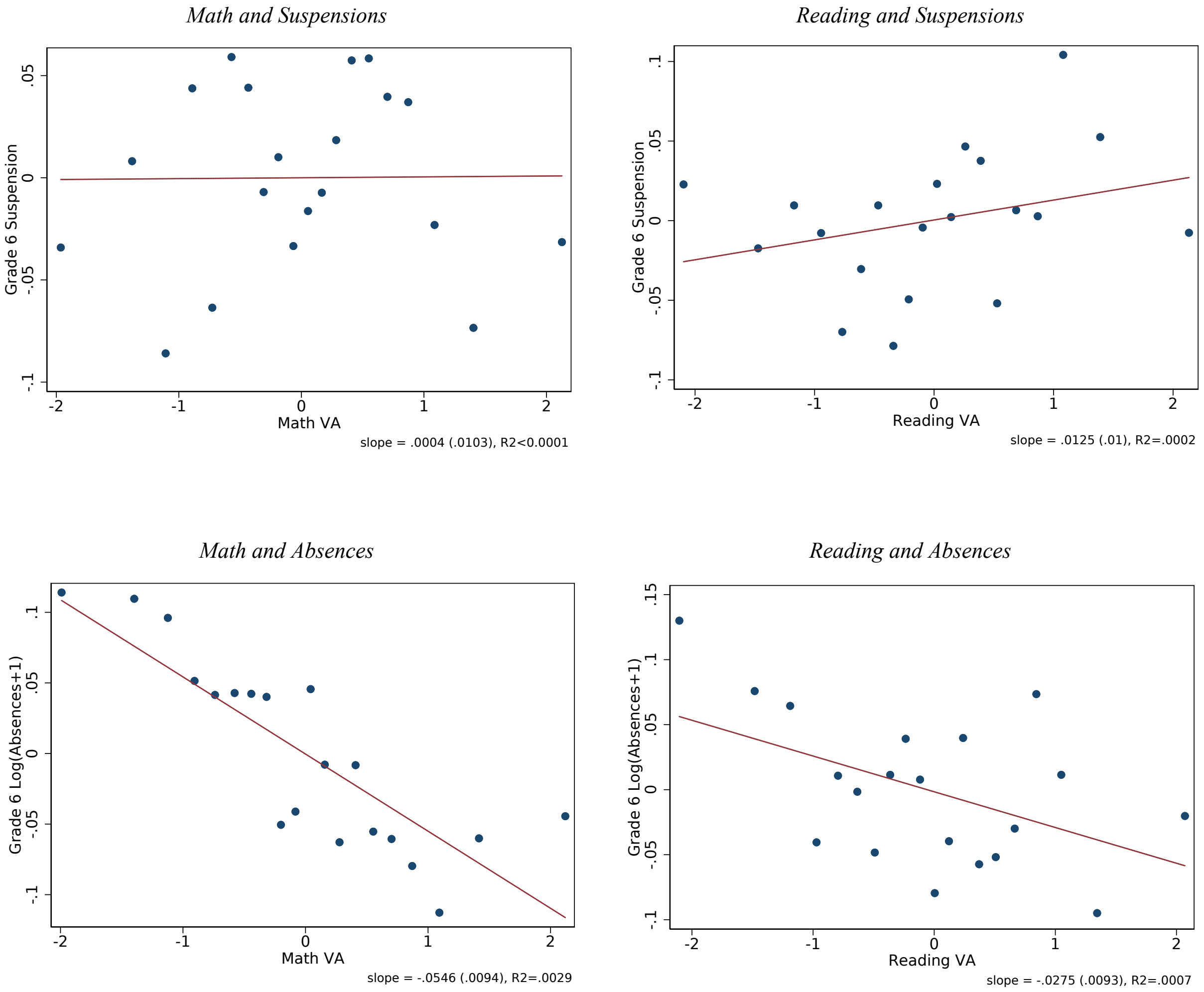

A. By Gender

B. By Race

C. By Economic Background

Notes: This figure displays the associations between drug charges and test scores for different subpopulations in grades 3 through 8. Points are slope coefficients from regressions of an indicator for a drug charge by age 20 on test scores and control variables. The control variables are the same as the controls used in the regressions from Figure 2. Panel A plots the coefficients for boys and girls, Panel B for black students and white students, and Panel C for students who are economically disadvantaged and non-disadvantaged. Each regression is estimated separately by grade. Error bars display 95% confidence intervals.

Figure 1.11: Binned Scatter Plots of Value-Added Measures



Math and Suspensions

slope = .0004 (.0103), R2<0.0001

Reading and Suspensions

slope = .0125 (.01), R2=.0002

Math and Absences

slope = -.0546 (.0094), R2=.0029

Reading and Absences

slope = -.0275 (.0093), R2=.0007

Notes: This figure displays the binned scatter plots between our cognitive and non-cognitive value-added measures of teacher quality. Our cognitive measures estimate VA using math and reading test scores, and our non-cognitive VA measures measure changes in grade 6 suspensions and absences. These measures were estimated following the procedure described in Section 1.4 of the paper, and then standardized to be mean 0 and standard deviation 1 within the distribution of teachers. The binned scatter plots display the mean cognitive and non-cognitive VA within each vingtile of the cognitive VA. The slopes represent the change in non-cognitive VA (in standard deviation units) associated with a one standard deviation change in test-score VA.

1.14 Appendix: Additional Tables

Table 1.7: Summary Statistics for Middle School Sample

	Value-Added Sample			Offenders' Sample at Age 20	
	Grade 6	Grade 7	Grade 8	Any	Felony
Female	0.493	0.494	0.495	0.357	0.174
Black	0.260	0.264	0.267	0.383	0.507
Economically Disadvantaged	0.515	0.500	0.483	0.558	0.681
Special Ed	0.134	0.123	0.116	0.151	0.206
LEP	0.049	0.046	0.044	0.026	0.025
Parents Attended College	0.419	0.421	0.427	0.330	0.228
Suspended	0.103	0.119	0.122	0.203	0.318
End-of-Year Math Score	0.028	0.034	0.040	-0.274	-0.554
End-of-Year English Score	0.022	0.028	0.033	-0.269	-0.571
Criminal Charge by Age 20	0.210	0.214	0.216	–	–
Felony Charge by Age 20	0.052	0.050	0.048	–	–
Criminal Conviction by Age 20	0.044	0.043	0.041	–	–
Felony Conviction by Age 20	0.019	0.018	0.017	–	–
N	972451	979691	976271	70371	16573

Notes: This table displays descriptive statistics for North Carolina middle school students matched to criminal records. The first three columns report mean demographics of sixth, seventh and eighth grade students restricted to individuals with observed math and reading test scores in the relevant grade. This sample is restricted to students in the 2007 cohorts and beyond, as this population can be matched to middle school teachers using course membership files. The first two columns report mean demographics of sixth, seventh, and eighth grade students, restricted to individuals with observed math and reading test scores in the given grade. Column (4) reports mean characteristics for our sample of individuals who have received a criminal charge by age 20. Column 5 reports mean characteristics for individuals with a felony charge by age 20.

Table 1.8: Estimating Forecast Bias in Value-Added Estimates

VA Measure		Standard Deviation in Outcome (1)	 (2)
Cognitive	Math Scores	1.165	1.049
		(0.016)	(0.011)
	Reading Scores	1.214	1.038
		(0.023)	(0.014)
Non-cognitive	Contemporaneous Suspensions	1.091	1.099
		(0.037)	(0.035)
	Future Suspensions	0.991	0.984
		(0.048)	(0.045)
	Contemporaneous log(Absences+1)	1.132	1.111
		(0.029)	(0.027)
	Future log(Absences+1)	1.202	1.181
		(0.030)	(0.027)
Demographic Controls			X

Notes: This table displays estimates of λ_1 in Equation (1.8) to quantify forecast bias in our value-added estimates. The row specifies the teacher VA measure used as our quality measure. For each measure of teacher quality, these estimates were constructed by first aggregating the skill measure and the teacher VA measure to the school-grade year level, and then by regressing first differences of the former on the first differences of the latter. Column (1) displays these estimates without including for differences in our control measure ξ. Column (2) includes changes in this measure. Estimates of ξ were constructed by estimating (1), replacing teacher dummies with teacher VA, and then aggregating the non-VA terms to the school-grade-year level.

Table 1.9: Impacts of Value-Added on Criminal Activity by Race

	Black Students				White Students			
	Charges at Age 20		Convictions at Age 20		Charges at Age 20		Convictions at Age 20	
Value-Added	Any	Felony	Any	Felony	Any	Felony	Any	Felony
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Math Scores	-0.223	-0.025	-0.199	-0.116	0.041	-0.040	-0.042	-0.017
	(0.099)	(0.064)	(0.061)	(0.044)	(0.062)	(0.030)	(0.029)	(0.017)
Reading Scores	-0.219	0.000	-0.086	-0.056	-0.018	-0.058	-0.096	-0.024
	(0.098)	(0.063)	(0.061)	(0.044)	(0.064)	(0.031)	(0.030)	(0.017)
Future	0.493	0.336	0.433	0.201	0.300	0.147	0.098	0.061
Suspensions	(0.140)	(0.090)	(0.085)	(0.061)	(0.109)	(0.052)	(0.050)	(0.029)
Future	0.665	0.492	0.334	0.318	0.295	0.113	0.002	0.049
log(Absences+1)	(0.202)	(0.128)	(0.119)	(0.086)	(0.133)	(0.063)	(0.059)	(0.035)

Notes: This table reports estimates of relationships between crime outcomes and teacher cognitive and non-cognitive value-added, segmenting the population by the race of the student. Estimates come from regressions of a crime indicator on posterior mean predictions of teacher value-added (Equation (1.8)). Coefficients are multiplied by 100. Value-added measures are standardized so that the reported estimates can be interpreted as the impact of a one-standard deviation change in value-added on the likelihood of committing a criminal offense by age 20, in percentage points. Columns (1) through (4) report estimates for black students. Column (1) reports the teacher impacts on any criminal charges, Column (2) reports the impacts on felony charges, Column (3) on any convictions, and Column (4) on felony convictions. Columns (5) through (8) report the corresponding estimates for white students. Standard errors are reported in parentheses and are clustered at the teacher level.

Table 1.10: Impacts of Value-Added on Criminal Activity by Family Economic Status

	Economically Disadvantaged				Not Economically Disadvantaged			
	Charges at Age 20		Convictions at Age 20		Charges at Age 20		Convictions at Age 20	
Value-Added	Any	Felony	Any	Felony	Any	Felony	Any	Felony
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Math Scores	-0.102	-0.004	-0.142	-0.071	0.018	-0.055	-0.054	-0.025
	(0.078)	(0.048)	(0.048)	(0.033)	(0.063)	(0.028)	(0.026)	(0.016)
Reading Scores	-0.234	-0.038	-0.136	-0.060	0.005	-0.037	-0.057	-0.014
	(0.079)	(0.049)	(0.048)	(0.033)	(0.065)	(0.029)	(0.027)	(0.016)
Future	0.403	0.304	0.365	0.200	0.395	0.161	0.107	0.044
Suspensions	(0.115)	(0.071)	(0.069)	(0.047)	(0.110)	(0.049)	(0.045)	(0.027)
Future	0.614	0.469	0.298	0.306	0.304	0.033	-0.018	-0.017
log(Absences+1)	(0.151)	(0.091)	(0.087)	(0.059)	(0.136)	(0.060)	(0.054)	(0.032)

Notes: This table reports estimates of relationships between crime outcomes and teacher cognitive and non-cognitive value-added, segmenting the population by the economic status of the student's family. Estimates come from regressions of a crime indicator on posterior mean predictions of teacher value-added (Equation (1.8)). Coefficients are multiplied by 100. Value-added measures are standardized so that the reported estimates can be interpreted as the impact of a one-standard deviation change in value-added on the likelihood of committing a criminal offense by age 20, in percentage points. Columns (1) through (4) report estimates for economically disadvantaged students. Column (1) reports the teacher impacts on any criminal charges, Column (2) reports the impacts on felony charges, Column (3) on any convictions, and Column (4) on felony convictions. Columns (5) through (8) report the corresponding estimates for non-disadvantaged students. Standard errors are reported in parentheses and are clustered at the teacher level.

Table 1.11: Pooled Impacts of Value-Added on Criminal Activity by Age of Outcome

		Math Scores				Reading Scores			
		Charges at Age 20		Convictions at Age 20		Charges at Age 20		Convictions at Age 20	
		Any	Felony	Any	Felony	Any	Felony	Any	Felony
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Cognitive Value-Added	17	-0.013 (0.023)	0.008 (0.010)	-0.018 (0.010)	-0.002 (0.006)	-0.034 (0.024)	0.005 (0.010)	-0.011 (0.010)	-0.006 (0.006)
	20	-0.028 (0.049)	-0.019 (0.027)	-0.084 (0.026)	-0.039 (0.017)	-0.104 (0.050)	-0.033 (0.027)	-0.087 (0.026)	-0.032 (0.017)
	23	-0.037 (0.074)	-0.049 (0.046)	-0.096 (0.045)	-0.099 (0.031)	-0.099 (0.075)	-0.067 (0.046)	-0.126 (0.045)	-0.085 (0.031)

		Future Suspensions				Future log(Absences+1)			
		Charges at Age 20		Convictions at Age 20		Charges at Age 20		Convictions at Age 20	
		Any	Felony	Any	Felony	Any	Felony	Any	Felony
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel B: Non-Cognitive Value-Added	17	0.185 (0.032)	0.051 (0.014)	0.059 (0.014)	0.024 (0.008)	0.194 (0.032)	0.090 (0.014)	0.071 (0.013)	0.024 (0.008)
	20	0.417 (0.079)	0.243 (0.043)	0.247 (0.041)	0.130 (0.027)	0.480 (0.102)	0.265 (0.055)	0.149 (0.051)	0.153 (0.034)
	23	0.424 (0.145)	0.277 (0.092)	0.283 (0.089)	0.220 (0.062)	-	-	-	-

Notes: This table reports estimates of relationships between crime outcomes at various ages and teacher cognitive and non-cognitive value-added. Estimates come from regressions of a crime indicator on posterior mean predictions of teacher value-added (Equation (1.8)). Coefficients are multiplied by 100. Value-added measures are standardized so that the reported estimates can be interpreted as the impact of a one-standard deviation change in value-added on the likelihood of committing a criminal offense by the given ages, 17, 20, and 23, in percentage points. The one exception is the effect of absences VA on criminal outcomes at age 23, as the attendance records did not begin early enough to connect absences to crime for 23-year-olds. Panel A uses cognitive measures of value-added (VA); Columns (1) through (4) use math VA and Columns (5) through (8) use reading VA. Panel B uses non-cognitive measures of value-added (VA); Columns (1) through (4) use 6th grade suspension VA and Columns (5) through (8) use 6th grade log(absences+1) VA. Standard errors are reported in parentheses and are clustered at the teacher level.

Table 1.12: Impacts of Value-Added on Criminal Activity by Teacher Grade

	Grade 4 Teachers				Grade 5 Teachers			
	Charges at Age 20		Convictions at Age 20		Charges at Age 20		Convictions at Age 20	
Value-Added	Any	Felony	Any	Felony	Any	Felony	Any	Felony
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Math Scores	-0.131 (0.069)	-0.079 (0.037)	-0.112 (0.036)	-0.067 (0.024)	0.078 (0.072)	0.044 (0.039)	-0.054 (0.037)	-0.010 (0.025)
Reading Scores	-0.204 (0.070)	-0.096 (0.038)	-0.086 (0.037)	-0.051 (0.024)	0.002 (0.073)	0.041 (0.040)	-0.084 (0.038)	-0.009 (0.025)
Future Suspensions	0.328 (0.128)	0.211 (0.069)	0.246 (0.066)	0.109 (0.044)	0.478 (0.101)	0.268 (0.055)	0.255 (0.052)	0.145 (0.034)
Future log(Absences+1)	0.709 (0.181)	0.328 (0.099)	0.295 (0.091)	0.225 (0.061)	0.370 (0.123)	0.233 (0.066)	0.082 (0.062)	0.115 (0.041)

Notes: This table reports estimates of relationships between crime outcomes and teacher cognitive and non-cognitive value-added, segmenting the population by the grade. Estimates come from regressions of a crime indicator on posterior mean predictions of teacher value-added (Equation (1.8)). Coefficients are multiplied by 100. Value-added measures are standardized so that the reported estimates can be interpreted as the impact of a one-standard deviation change in value-added on the likelihood of committing a criminal offense by age 20, in percentage points. Columns (1) through (4) report estimates for fourth grade teachers. Column (1) reports the teacher impacts on any criminal charges, Column (2) reports the impacts on felony charges, Column (3) on any convictions, and Column (4) on felony convictions. Columns (5) through (8) report the corresponding estimates for fifth grade teachers students. Standard errors are reported in parentheses and are clustered at the teacher level.

Table 1.13: Value-Added Measures Using Combinations of Elementary School Teacher Effects

	Cumulative Effects				Average Effects			
	Charges at Age 20		Convictions at Age 20		Charges at Age 20		Convictions at Age 20	
Value-Added	Any	Felony	Any	Felony	Any	Felony	Any	Felony
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Math Scores	0.016	-0.008	-0.052	-0.024	0.009	-0.009	-0.074	-0.029
	(0.064)	(0.032)	(0.034)	(0.020)	(0.055)	(0.028)	(0.030)	(0.018)
Reading Scores	-0.037	-0.009	-0.036	-0.026	-0.036	-0.017	-0.070	-0.023
	(0.065)	(0.033)	(0.034)	(0.021)	(0.056)	(0.028)	(0.029)	(0.018)
Future	0.429	0.202	0.264	0.119	0.370	0.200	0.206	0.105
Suspensions	(0.120)	(0.062)	(0.069)	(0.040)	(0.085)	(0.046)	(0.049)	(0.029)
Future	0.363	0.113	0.069	0.086	0.386	0.212	0.135	0.125
log(Absences+1)	(0.133)	(0.067)	(0.066)	(0.042)	(0.099)	(0.051)	(0.050)	(0.032)

Notes: This table reports estimates of relationships between crime outcomes and cognitive and non-cognitive VA. These estimates use the student as the unit of observation, rather than the student-year observations presented in all other tables, and show the effects of a students' total elementary school teachers' VA on crime (rather than just the effect of an individual teacher in a given year). Estimates come from regressions of a crime indicator on posterior mean predictions of teacher value-added (Equation (1.8)). Coefficients are multiplied by 100. Value-added measures are standardized so that the reported estimates can be interpreted as the impact of a one-standard deviation change in value-added on the likelihood of committing a criminal offense by age 20, in percentage points. Columns (1) through (4) use the sum of the fourth and fifth grade value-added scores as the student's cumulative value-added. Column (1) reports the teacher impacts on any criminal charges, Column (2) reports the impacts on felony charges, Column (3) on any convictions, and Column (4) on felony convictions. Columns (5) through (8) show the corresponding estimates using the average VA across elementary school teachers. These averages omit missing VA, meaning that if a given student only had one observed teacher VA, then that teacher's VA is used as the student's average. Standard errors are reported in parentheses and are clustered at the teacher level.

Table 1.14: Impacts of Value-Added on Crime Outcomes Using Alternative Shrinkage Procedures

	Unconditional Shrinkage Without Controls				No Shrinkage			
	Charges at Age 20		Convictions at Age 20		Charges at Age 20		Convictions at Age 20	
Value-Added	Any	Felony	Any	Felony	Any	Felony	Any	Felony
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Math Scores	-0.026	-0.019	-0.083	-0.038	-0.025	-0.018	-0.082	-0.038
	(0.049)	(0.027)	(0.026)	(0.017)	(0.050)	(0.027)	(0.026)	(0.017)
Reading Scores	-0.087	-0.029	-0.081	-0.029	-0.091	-0.032	-0.076	-0.034
	(0.049)	(0.027)	(0.026)	(0.017)	(0.052)	(0.028)	(0.027)	(0.018)
Future	0.367	0.201	0.205	0.108	0.301	0.195	0.198	0.098
Suspensions	(0.072)	(0.039)	(0.037)	(0.024)	(0.071)	(0.039)	(0.037)	(0.024)
Future	0.443	0.217	0.120	0.139	0.364	0.193	0.117	0.147
log(Absences+1)	(0.097)	(0.052)	(0.049)	(0.032)	(0.096)	(0.052)	(0.048)	(0.032)

Notes: This table reports estimates of relationships between crime outcomes and teacher cognitive and non-cognitive value-added, using alternative empirical Bayesian methods to reduce forecast bias of the estimated teacher effects. Estimates come from regressions of a crime indicator on posterior mean predictions of teacher value-added (Equation (1.8)). Coefficients are multiplied by 100. Value-added measures are standardized so that the reported estimates can be interpreted as the impact of a one-standard deviation change in value-added on the likelihood of committing a criminal offense by age 20, in percentage points. Columns (1) through (4) use a shrinkage without control variables - all teachers are assumed to be drawn from the same distribution of quality with a fixed mean and variance. Column (1) reports the teacher impacts on any criminal charges, Column (2) reports the impacts on felony charges, Column (3) on any convictions, and Column (4) on felony convictions. Columns (5) through (8) report the corresponding estimates without employing any shrinkage - these estimates simply use the standardized fixed effects estimates of (1) as our VA measures. Standard errors are clustered at the teacher level.

Table 1.15: Impacts of Value-Added on Criminal Activity for Middle School Teachers

	Charges at Age 20		Convictions at Age 20	
VA Measure	Any	Felony	Any	Felony
	(1)	(2)	(3)	(4)
Math Scores	0.008 (0.033)	-0.031 (0.015)	-0.034 (0.014)	-0.011 (0.009)
Reading Scores	-0.286 (0.037)	-0.113 (0.017)	-0.126 (0.016)	-0.031 (0.010)
Suspensions	0.301 (0.033)	0.113 (0.015)	0.151 (0.014)	0.088 (0.009)
Future Suspensions	0.319 (0.033)	0.128 (0.015)	0.158 (0.014)	0.094 (0.009)
Contemporaneous log(Absences+1)	0.391 (0.037)	0.106 (0.017)	0.096 (0.016)	0.038 (0.010)
Future log(Absences+1)	0.253 (0.032)	0.051 (0.015)	0.012 (0.014)	0.036 (0.009)

Notes: This table reports estimates of relationships between crime outcomes and middle school teacher cognitive and non-cognitive value-added. Estimates come from regressions of a crime indicator on posterior mean predictions of teacher value-added (Equation (1.8)). Coefficients are multiplied by 100. Value-added measures are standardized so that the reported estimates can be interpreted as the impact of a one-standard deviation change in value-added on the likelihood of committing a criminal offense by age 20, in percentage points. Column (1) reports the teacher impacts on any criminal charges, Column (2) reports the impacts on felony charges, Column (3) on any convictions, and Column (4) on felony convictions. Standard errors are reported in parentheses and are clustered at the teacher level.

Table 1.16: Impacts of Value-Added on Criminal Activity at Age 18 for North Carolina High School Students

	In Grade 9				In Grade 12			
	Charges at Age 20		Convictions at Age 20		Charges at Age 20		Convictions at Age 20	
Value-Added	Any	Felony	Any	Felony	Any	Felony	Any	Felony
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Math Scores	-0.223	-0.025	-0.199	-0.116	0.041	-0.040	-0.042	-0.017
	(0.099)	(0.064)	(0.061)	(0.044)	(0.062)	(0.030)	(0.029)	(0.017)
Reading	-0.008	0.005	-0.013	-0.004	-0.055	0.003	-0.037	-0.015
	(0.035)	(0.012)	(0.011)	(0.005)	(0.037)	(0.017)	(0.017)	(0.010)
Future	0.154	0.020	0.049	0.003	0.257	0.096	0.103	0.052
Suspensions	(0.050)	(0.017)	(0.015)	(0.007)	(0.052)	(0.024)	(0.024)	(0.014)
Future	0.218	0.049	0.016	-0.002	0.382	0.155	0.114	0.042
log(Absences+1)	(0.052)	(0.017)	(0.015)	(0.007)	(0.055)	(0.025)	(0.024)	(0.015)

Notes: This table reports estimates of relationships between crime outcomes and teacher cognitive and non-cognitive value-added, restricting the sample to students who appear in the education records as high school to avoid interstate attrition from the criminal records sample. Estimates come from regressions of a crime indicator on posterior mean predictions of teacher value-added (Equation (1.8)). Coefficients are multiplied by 100. Value-added measures are standardized so that the reported estimates can be interpreted as the impact of a one-standard deviation change in value-added on the likelihood of committing a criminal offense by age 18, in percentage points. Columns (1) through (4) report estimates for students who appear as ninth graders in the education records. Column (1) reports the teacher impacts on any criminal charges, Column (2) reports the impacts on felony charges, Column (3) on any convictions, and Column (4) on felony convictions. Columns (5) through (8) report the corresponding estimates for students who appear as twelfth graders in the education records. As the majority of students will turn 18 by the completion of grade 12, the estimates in columns (5) through (8) should be unaffected by migration. Standard errors are clustered at the teacher level.

Table 1.17: Impacts of Value-Added on Criminal Activity at Age 20 for North Carolina High School Students

	In Grade 9				In Grade 12			
	Charges at Age 20		Convictions at Age 20		Charges at Age 20		Convictions at Age 20	
VA	Any	Felony	Any	Felony	Any	Felony	Any	Felony
Math Score	0.002 (0.057)	-0.042 (0.024)	-0.060 (0.021)	-0.017 (0.011)	-0.060 (0.054)	-0.029 (0.029)	-0.102 (0.028)	-0.040 (0.018)
Reading Scores	-0.007 (0.058)	-0.003 (0.024)	-0.051 (0.022)	-0.002 (0.012)	-0.114 (0.055)	-0.043 (0.029)	-0.103 (0.028)	-0.033 (0.018)
Future Suspensions	0.411 (0.092)	0.136 (0.038)	0.161 (0.033)	0.039 (0.018)	0.424 (0.086)	0.222 (0.046)	0.230 (0.044)	0.115 (0.028)
Future log(Absences+1)	0.257 (0.116)	0.071 (0.049)	-0.009 (0.042)	0.004 (0.022)	0.550 (0.112)	0.288 (0.060)	0.138 (0.055)	0.150 (0.036)

Notes: This table reports estimates of relationships between crime outcomes and teacher cognitive and non-cognitive value-added, restricting the sample to students who appear in the education records as high school to avoid interstate attrition from the criminal records sample. Estimates come from regressions of a crime indicator on posterior mean predictions of teacher value-added (Equation (1.8)). Coefficients are multiplied by 100. Value-added measures are standardized so that the reported estimates can be interpreted as the impact of a one-standard deviation change in value-added on the likelihood of committing a criminal offense by age 20, in percentage points. Columns (1) through (4) report estimates for students who appear as ninth graders in the education records. Column (1) reports the teacher impacts on any criminal charges, Column (2) reports the impacts on felony charges, Column (3) on any convictions, and Column (4) on felony convictions. Columns (5) through (8) report the corresponding estimates for students who appear as twelfth graders in the education records. Standard errors are clustered at the teacher level.

Chapter 2

Can Racial Gaps in Offending Be Explained?

2.1 Introduction

Racial disparities in the United States criminal justice system are large and have grown substantially in the past 50 years. Between 1970 and 2000, while the fraction of white men who were incarcerated remained roughly constant at about 1%, the fraction of black men in jail or prison grew from 3% to 8% (Raphael, 2006). This gap is particularly large for young men without a high school degree – 40% of black male dropouts under age 30 were incarcerated at the turn of the century, quadruple the fraction of their white counterparts (Western and Pettit, 2002).

Explanations of these massive differences in crime outcomes across races have been studied extensively, often to investigate if these gaps are attributable to differences in black and white populations that affect relative criminal propensities, or if these differences are due to discrimination in the criminal justice system. Many environmental factors have been shown to affect the magnitude of the crime gap, including local levels of segregation (Shihadeh and Flynn, 1996) and inequality (Harer and Steffensmeier, 1992), "white-flight" (Liska and Bellair, 1995), and educational attainment (Lochner and Moretti, 2004). Past studies have also documented differential treatment of blacks and whites at various stages of the criminal justice process, including when police use force (Fryer, forthcoming), when booking juvenile offenders (Raphael and Rozo, 2019), and when making bail decisions (Arnold, Dobbie, and Yang, 2018).

We present new evidence on the black-white gap in crime by considering the explanatory power of another factor: cognitive skill. Large gaps in academic achievement measures between black and white students evolve at very early ages and remain present throughout early years of schooling (Jencks and Phillips, 1998; Fryer and Levitt, 2006). This "skill-gap" has been shown to explain a substantial fraction of racial differences in other long-run outcomes including wages, employment, and health (Neal and Johnson, 1996; Fryer, 2011).

Average differences in black and white achievement may also influence differences in crime rates across races. Moreover, if cognitive skills are rewarded differently for black and white students, returns to skills could impact their relative rates of offending.

Fryer (2011) provides, to our knowledge, the first direct evidence that differences in educational achievement between blacks and whites can partially explain the large differences in incarceration rates using the National Longitudinal Survey of Youth 1997 (NLSY97). We extend Fryer's analysis using administrative records. We utilize a new dataset linking North Carolina public school student records to all future court records in the state. This linkage provides detailed information about the student's home and educational environment, academic ability at early ages using test scores, and all future criminal charges and convictions that occur in early adulthood.

We begin our analysis by following Neal and Johnson (1996), who explore the role of premarket factors in explaining the black-white wage gap. Our rich data allow us to explore how observable differences across black and white populations in many characteristics, including differences in socio-economic status, schools, neighborhoods, and educational achievement, can explain the differences in black-white offending rates. We primarily focus on the explanatory power of test scores. We also explore the relevance of educational achievement in explaining racial crime gaps for subpopulations that are at greater risk, namely boys and students from economically disadvantaged backgrounds. We focus on the extensive margin, specifically differences in black and white students' rates of engagement in any criminal activity in early adulthood.

Next, to fully characterize the explanatory power of achievement on differences in offending rates, we perform a Blinder-Oaxaca (Blinder, 1973; Oaxaca, 1973) decomposition of the crime gap. This analysis quantifies how much of the gap in offending can be explained by (1) differences in environment for black and white students, (2) differences in average achievement levels between black and white students, and (3) differences in returns to skill by race.

Our analysis reveals three sets of findings. First, we find that variation in test scores explains a significant fraction of the differences in black and white crime rates, even after controlling for a rich set of other covariates. We find that test scores prior to high school can explain between a quarter to a half of the black-white gap in charge rates for any criminal offense by the age of 20. Moreover, the residual gap in overall offending rates becomes statistically insignificant after controlling for test scores (in addition to our other characteristics). This pattern appears to be driven by test scores explaining differences in charge rates for less severe offenses, which constitute the vast majority of our sample. While test scores still account for a substantial fraction of the differences in black and white felony offending rates, controlling for them does not eliminate this residual gap.

Second, we document substantial differences between black and white students' returns to skill. Black students experience a much greater return to test scores through crime reduction than white students. These differential returns explain a substantial fraction – between 10% and 20% – of the raw crime gap. This is especially true for more severe offenses – when we focus on the black-white gap in felony offending rates, we find that differences in returns to test scores account for nearly the same fraction of the raw gap as the differences in average test scores.

Third, we find significant heterogeneity in the explanatory power test scores on the crime gap, mainly by socio-economic status. Test scores are particularly powerful in explaining differences in overall charge rates of black and white students from economically disadvantaged backgrounds – they explain over 2/3 of the raw gap in this subpopulation, with a large fraction (30%) being attributable to differential returns to skill. Test scores are able to explain less of the gap in felony offending rates for this population, although differential returns account for a larger fraction of the gap than differences in average test scores. On the other hand, when restricting to students from non-disadvantaged backgrounds, black students are less likely to be charged with any offense at an early age than their white counterparts (although black students are more likely to commit felonies), conntrolling for test scores widens this crime gap.

We conclude our analysis by discussing several possible explanations for these empirical facts. We preview several possible theories for how differences in test score levels and returns across races affect the crime gap, focusing primarily how cognitive skill may incentivize one's decision to engage in criminal activity as opposed to legal employment (Freeman, 1999), and how these incentives may vary by group. We emphasize that while the these empirical facts are interesting, this study is not causal, and testing our theories is beyond the scope of this research. We hope to expand this analysis in future work.

The rest of the paper is as follows. Section 2.2 provides institutional context and summarizes our data. Section 2.3 provides descriptive evidence of the explanatory power of test scores, both through levels and returns, on the black-white crime gap, and Section 2.4 formally decomposes the gap. Section 2.5 discusses possible explanations of our results and concludes.

2.2 Data

We conduct our analysis using a unique dataset that merges public school administrative education records in North Carolina with all criminal charge and conviction records in the state. The education records were provided by the North Carolina Education Research Data Center and include detailed records of all public school students, including demographic information, attendance records, coarse address information, and test scores. In North Carolina, all students in grades 3 though 8 take standardized math and reading tests at the end of the year.

We then linked these education records to two sets of criminal records: charges and convictions. Our charge records come form the North Carolina District and Superior Courts, and comprise all formal charges for any misdemeanor or felony offense occurring in the state between 2005 and 2015. We also have sentencing records provided by the Department of Public Safety for all convictions since between 1970 and 2015 that required a period of mandatory supervision as part of the sentence (either probation or incarceration). For the majority of this paper, however, we will use the charges data, as the patterns we see in the convictions data very closely follow the patterns from the felony charges.

We restrict our charges sample to all North Carolina public school students born between 1989 and 1995, as these students' criminal charge history between ages 16 and 20 would appear in our data.[1] When focusing on convictions, we expand the sample to include all individuals born no later than 1995 that attend North Carolina public schools after the 1992-1993 school year. We further restrict our sample to only black and white students to focus solely the black-white gap in offending rates by age 20.

Column (1) in Table 2.1 describes our sample used to analyze criminal charges. (The sample used to analyze our conviction records can be found the appendix.) About one third of our sample is black and two thirds are white. We see the majority of our students come from economically disadvantaged backgrounds, and about half of parents report ever attending college.

Columns (2) and (3) summarize our samples for blacks and whites separately. These samples of students are very observably different. Over 80% of the black students in our sample are from economically disadvantaged backgrounds and are less likely to graduate high school. On average, black students score about three quarters of a standard deviation below whites, and this gap is persistent and roughly constant across all grades that we observe. The magnitude of the test score gap and its evolution over grades is consistent with work by Fryer and Levitt (2006, 2013), who show that these test score gaps emerge and evolve in even earlier grades, and then persist in later adolescence.

The final three columns describe our population of offenders, or anyone who was charged with an offense between ages 16 and 20. The sample is heavily negatively selected - these offenders are much more likely to be from disadvantaged backgrounds, are less likely to have parents who attended college, and have worse test scores than the general population by about a quarter of a standard deviation. A much larger fraction of this population is black than in the general population. In the appendix, we see these patterns are exacerbated for more serious offenses.

2.3 Observed Relationship Between Test Scores and Crime

Observed Correlations

We begin with the raw correlations between test scores and crime rates, and show how their observed correlations vary by race. Figure 2.1 displays binned scatter plots of test scores and crime rates by age 20. These figures were made by splitting black and white students into vingtiles of their race's respective test score distributions, and these graphs plot the average

[1] All adolescents that are at least 15 years of age are charged as adults. For this analysis, the minimum offending age will be age 16, since in practice, very few 15-year-olds are charged.

offending rate and test score in each vingtile. This figure shows four of these binscatters, plotting average rates of initial criminal charges before age 20 vs third and eighth grade math and reading scores.

Three patterns emerge from these binned scatter plots. First is the general, approximately linear relationships that exist across all races and for all test scores. The grade 8 test score graphs are steeper than those for grade 3; as the scores are normalized across years to have standard deviation 1, this means that the variation in test scores in grade 8 explains a greater portion of the variation in crime rates. This pattern is consistent with a measurement error story, in which later tests are more accurate measures of achievement (Bond and Lang, 2018). We see this pattern holds for both reading and math test scores.[2]

Second is how the offending rates between black and white students compare throughout the test score distribution. We see that black students are charged more frequently than white for all students who score less than half a standard deviation above the mean, or approximately the 75th percentile of the overall test score distribution. Above this threshold, we see no substantial difference between charge rates of white and blacks - if anything, black students with high test scores are charged less frequently than whites.

Third, the crime-test score gradients are substantially different for black and white students. Black students experience a higher return to test scores in the form of reduced crime. This result is in contrast black-white gaps in other long-run outcomes. Returns to skill in the form of wages are very similar for blacks and whites (Neal and Johnson, 1996).

Figure 2.2 further parses these relationships to highlight the explanatory power of test scores for different types of crimes, namely misdemeanor charges, felony charges, and convictions. We see similar patterns for all types of crimes. Notably, blacks experience even stronger relative returns to higher test scores for more serious crimes. In the appendix, we also look at differences in returns to test scores for different classifications of crimes, namely drug crimes, assaults, and property crimes (e.g. larceny). We see a similar pattern, where black students experience greater returns to skill more serious offenses.[3]

Figure 2.3 highlights the importance of the socio-economic status of the student. We plot binned scatter plots for misdemeanors and felonies for both economically disadvantaged and non-disadvantaged kids. For non-disadvantaged students, black students are less likely to be charged with lesser crimes than their white counterparts, and there is no sizable difference in returns to skill across races for these offenses. However, we still see higher rates of felony charges for this population, and greater returns to test scores. For students from worse economic backgrounds, black students are charged more frequently and experience greater returns for both misdemeanors and felonies. The black-white difference in returns to skill for felony rates is particularly large for economically disadvantaged students. We display similar binscatters with test scores and crime outcomes for other subgroups in the appendix. The

[2] The other figures presented in this chapter only show the relationship between crime and math scores, but very similar patterns exist between crime and reading scores.

[3] Over 90% of the property crimes in our data are felonies, and over 90% of the drug crimes are misdemeanors. The overwhelming majority of assaults are also misdemeanors, but convictions for assault often carry strong consequences, requiring either probation or incarceration as part of the sentence.

relative rates of offending and differences in returns to skill are even larger for economically disadvantaged boys.

These binned scatter plots highlight the main patterns we find in this paper: test scores can partially explain the differences in the offending rates of blacks and whites, and the returns to test scores are larger for black students, particularly when considering serious offenses of students from economically disadvantaged backgrounds. However, test scores are related to (and influenced by) a multitude of other factors, and these relationships between criminal behavior and test scores may reflect how other characteristics of the student affect crime. Using our other rich demographic information, we will now further investigate the explanatory power of cognitive skill.

The Residual Black-White Crime Gap

We first test how much of the black-white gap in offending as young adults can be accounted for by test scores and other characteristics. We consider the following linear probability model:

$$Y_i = \alpha + \gamma \text{Black}_i + X_i' \beta + S_i' \Gamma + \epsilon_i \tag{2.1}$$

here, Y_i is a binary variable indicating if i committed a given criminal offense between the ages of 16 and 20, X_i are standardized test scores from 3rd to 8th grade, and S_i is a set of controls that can include demographic information about the student and their family (namely gender of the student, maximum parental level of education, and whether or not the student is enrolled in special education or is economically disadvantaged), cohort fixed effects, school fixed effects, and census tract fixed effects.

Table 2.2 describes the results of this regression with varying sets of controls. In Column (1), we display the naive gap, finding that black students are 8.8 percentage points more likely to be charged with a crime by age 20 whites, or about 40% of the average offending rate. However, once we control for all test scores between grades 3 through 8 in Column (2), we see that this gap shrinks by over 50%. In the appendix, we show that simply controlling for the most recent test scores has a similar if not greater impact on closing the black-white crime gap, suggesting that the most recent grades are the most relevant for explaining crime differences. In any case, we find that differences in these skills explain a large fraction of the gap.

We also use our rich set of controls to see how our test score explanation compares to the other observable characteristics in regards to explanation of the gap. Columns (3) through (5) are add progressively finer controls. Controlling for school and cohort reduces the residual gap by about 20%, and controlling for census tract in Column (4), reduces the gap by another 40%. After controlling for student and family characteristics in Column (5), the residual gap is statistically insignificant. Differences in observable characteristics between blacks and whites can explain the entire differences in offending rates by race.

While variation in test scores can explain a significant portion of the black-white gap in offending, it is possible that this could just be because test scores are highly correlated with some of our other controls (Gelbach, 2016). By first controlling for other characteristics and then adding test scores, we avoid attributing the explanatory power of test scores to other correlated explanatory variables (Fortin, Lemieux, and Firpo, 2011; Altonji, Bharadwaj, and Lange, 2012). Column (6) reports estimates of γ when including all controls except test scores.We see without controlling for test scores, the residual offending gap is about 2 percentage points, indicating that test scores explain at minimum about a quarter of the gap.

The Importance of Differential Returns to Achievement

In the previous sections, we have only considered a common returns to test scores across races. However, in our binned scatter plots, blacks experience a larger return to achievement, particularly for more severe crimes. Table 2.3 explores the relevance of this for all of our charge outcomes, and reports estimates of γ from Equation (2.1) when also controlling for an interaction between the test score and a black dummy variable. This allows for differential returns to skill across races.

We find that controlling for differential returns further explains the residual gap for felonies. Column (9) reports the residual gap after allowing for differential returns to skill. Adding returns closes about a third of the residual gap that existed when only allowing for a uniform return to skill for both blacks and whites. For the gap in any offending and less severe offending, we actually find that adding returns to skill (Columns (3) and (6), respectively) overaccounts for the racial gap in criminal activity, and that the differences in returns to test scores for blacks and white students are significantly different for all crime outcomes.

Robustness Checks

In the appendix, we compare our findings to national survey data. Similar to Fryer (2011), we use the National Longitudinal Sample of Youth 1997 (NLSY97), a nationally representative sample of individuals born in years 1980-1984, or about a decade before the students in our North Carolina sample. The test scores and crime outcomes are not perfectly comparable to our administrative records,[4] but we find similar patterns. The gaps in offending are comparable, and inclusion of test scores makes any residual differences in offending statistically insignificant, although these survey estimates of the gap are imprecise.

Additionally, we address the role of educational attainment, an important factor that may impact racial gaps in long-run outcomes. In our main specifications, we follow Neal and Johnson (1996) and do not control for the quantity of schooling, as the choice to obtain more

[4] The NLSY97 reports percentiles a weighted average of math and verbal scores from the Armed Services Vocational Aptitude Battery (ASVAB) test, similar to the Armed Forces Qualification Test (AFQT) score used by the department of defense. The crime outcomes are any arrests, which are comparable to our "any charges" outcome, and "any incarceration spells", which is more severe than any of our criminal outcomes.

education is endogenous. Lang and Manove (2011) show that if there is statistical discrimination in the labor market that mediated through educational attainment, then failing to control for schooling may overstate the test scores' ability to explain racial gaps. Similarly, if this labor market discrimination makes work more costly and crime more valuable for blacks, we may be overestimating how skill differences contribute to the black-white crime gap. Additionally, educational attainment may directly impact students' criminal behavior via incapacitation, as attending school limits students' ability to engage in criminal activity (Jacob and Lefgren, 2003).

In the Appendix Table 2.12, we control for maximum education attained, when we control for maximum level of educational achievement our main results remain unchanged. We find that after controlling for all observables, including educational attainment, there is still a statistically significant residual gap in offending rates of about 1 percentage point, or approximately 15% of the raw gap. Additionally, while controlling for test scores has less explanatory power when also controlling for schooling, we still find that they account for about 15-20% of the raw gap. Even when controlling for schooling attainment, we find that differences in skills helps explain the differences in black and white offending rates.

Heterogeneity

We compare these results for students from disadvantaged and non-disadvantaged backgrounds in Tables 2.4 and 2.5, respectively. For students from worse economic backgrounds, we find the same patterns as we see in the general population; if anything, differences in test scores explain a larger portion of the racial disparities here than they do on average. Column (2) of Table 2.4 shows that for poorer students, test score differences accounts for approximately half of the residual gap in black and white offending, after controlling for other characteristics. Moreover, after accounting for differences in returns to skill by race in Column 3, we are able to explain another half of the remaining gap. We see similar patterns for the importance of test scores for both misdemeanors and felonies.

However, in contrast to the other subgroups of interest, we find that non-disadvantaged black students are *less* likely to offend than their white counterparts, and that after controlling for test scores, this gap more than doubles in magnitude. This result is driven by differences in misdemeanor charge rates. We still see black students from better backgrounds commit a greater number of felonies, but after controlling for test scores, we can explain about 95% of the black-white gap in more severe offending for this subpopulation.

In appendix, we also look at if test scores have different explanatory power for different types of criminal offenses or if the relationship varies by gender. (Similar to the patterns depicted in the binned scatter plots, the results when using assault and property crime charges are very similar to the results when restricting to felony offenses, and the results for drug offenses are similar to the results for misdemeanors. However, observable characteristics other than test scores already explain the entire raw black-white gap in drug offenses.) We also focus on the gap in offending for men and women separately, and find that test scores are able to explain black-white offending gaps for both genders.

2.4 Decomposition of the Race Gap Using Test Scores

Estimation

To understand the extent to which our observables explain the black-white gap in criminal involvement, we conduct the following Oaxaca-Blinder (Blinder, 1973; Oaxaca, 1973) decomposition. We use the following econometric model that is estimated separately for blacks and whites:

$$Y_i = \alpha_{g(i)} + X_i'\beta_{g(i)} + S_i'\Gamma + \eta_i, \quad g(i) \in \{B, W\} \tag{2.2}$$

Here g represents the race of individual i. X_i and S_i are the same vectors of characteristics used in Equation (2.1). Importantly, this model does not allow for our demographics to vary in impact by race. Additionally, as we saw before, all of our controls in S are binary, meaning that by keeping the dependence of Y constant for all races, we are controlling for average offending rates for given demographic cells, and we are estimating the effects of test scores on the black-white crime gap within demographic groups.

Following Oaxaca (1973) and Blinder (1973), we can use (2.2) to decompose the observed raw gap, or average differences in offending across races, as follows:

$$\bar{Y}_B - \bar{Y}_W = \underbrace{(\bar{X}_B - \bar{X}_W)'\hat{\beta}_W}_{\text{Levels component}} + \underbrace{\bar{X}_B'(\hat{\beta}_B - \hat{\beta}_W)}_{\text{Returns component}} + \underbrace{(\bar{W}_B - \bar{W}_W)'\hat{\Gamma}}_{\text{Differences in controls}} + \underbrace{\hat{\alpha}_B - \hat{\alpha}_W}_{\text{Residual}} \tag{2.3}$$

Our assumption for uniform effects of S on crime for blacks and whites permits us to decompose the raw black-white gap in offending into four terms.[5] First, is a levels component in test scores. This term captures the portion of the gap explained by differences in average test scores between blacks and whites. This term reflects Neal and Johnson's (1996) "skill gap". Second is the differences in *returns* to crime from test scores across populations. This term captures how higher achievement is rewarded differentially by race, interpreted as a measure of discrimination by skill. The third component quantifies the contribution of average differences in non-test score characteristics S between blacks and whites to the crime gaps. Our final remaining term, the residual, can be interpreted as the difference offending rates for blacks and whites with average test scores, conditional on S.

Results

Table 2.6 displays our estimates of the breakdown from Equation (2.3) for several different outcomes (any charge, any misdemeanor charge, and any felony charge by age 20) and several different test scores (all tests, only grade 8 math and reading, and only grade 3 math and

[5] Under a more flexible model with Γ varying by race, we would not be able to separately identify the returns components from the residual gap (Jones, 1983; Oaxaca and Ransom, 1999).

reading). Unlike previous specifications, we restrict our sample to individuals with the given set of test scores.[6]

Columns (1) decomposes the black-white gap in overall offending rates at age 20 using all observed math and reading test scores in grades 3 through 8. Similar shown in Table 2.3, our observables can explain the entire raw gap in offending (in fact, we are over-explaining it). This breakdown reveals that approximately half of the raw gap can be explained by differences in test scores of the two populations. The other half can be explained by differences in the other observable characteristics of the students, including schools, neighborhoods, demographics, and family characteristics.

Columns (2) and (3) perform the same decomposition of the overall crime gap using only test scores from grades 8 and 3, respectively. The earliest observed test scores explain a smaller fraction of the raw gap – the level differences in third grade scores account for nearly one fifth of the raw gap, and returns component accounts for another tenth. However, when we use our most recent tests in grade 8, we find that test score levels and returns are able to explain about the same fraction of the gap, if not more, than when we use all tests. The increased explanatory power of later test scores is not due to a divergence in average test scores between blacks and whites over time, for the black-white test score gap stays roughly constant between grades 3 and 8 (as seen in Table 2.1).

This importance of the age at the test could be explained by a model of test scores noisily predicting ability, with less measurement error for more recent tests (Bond and Lang, 2018).[7] In Appendix Tables 2.16 and 2.17, we further explore the relationship between test scores at different ages and the black-white crime gap – we see the explanatory power of test scores for this crime gap are monotonically increasing with the grade at which the students take the test, which is consistent with this model.

Columns (4)-(6) displays the same breakdowns in the black-white crime gap for first-time misdemeanor charges by age 20. We see very similar patterns and scopes of explanatory power of test scores for these lesser offenses as we do when considering all offenses.[8] However, when we focus on our sample of felony offenders, the explanatory power of test scores changes substantially. Columns (7)-(9) display similar decompositions for felony charges. Contrary to offenders of lesser offenses, there is a sizable residual gap between blacks and whites that cannot be explained with our data – between 15% and 25% of the felony gap remains unexplained. We can still account for at least 40% of the raw gap with non-test score observables, which is similar to explanatory power of lesser offenses. However, the

[6] When estimating Equation (2.1), we assume all S and X are missing at random, added dummy variables for missing observations, and imputed the means for any missing values. We did the same thing here for S when estimating Equation (2.2).

[7] Another possibility is that the variance in actual achievement is growing over time, so our normalization is masking greater variation in test scores that explains the crime cap (Cascio and Staiger, 2012). However, if growing variation in test scores were driving this increase in explanatory power, we would expect to see a greater correlation in test scores between adjacent grades for higher grades vs lower grades. We see no evidence of such a pattern in our data.

[8] This is to be expected, since these populations are almost exactly the same. 95% of our first-time offenders have been charged with a misdemeanor.

explanatory power of the levels in test scores decreases to about 20% of the raw gap. The returns to test scores are a relatively larger factor in the explanatory power of test scores of the gap. When only using grade 8 test scores, about 40% of the gap explained by test scores, but nearly half of this fraction is due to differences in returns to skill for black and white students.

In addition to heterogeneity in types of crime and age of test, we see large differences by socioeconomic status. Table 2.7 decomposes the gaps in offending for all charges, misdemeanor charges, and felony charges for students from economically disadvantaged and non-disadvantaged households. For both subpopulations, there is a dramatic decrease in the explanatory power of our controls relative to the decompositions from Table 2.6, since economic status explains a large portion of the differences across races in offending rates.

Columns (2), (4), and (6) show the decomposition for students from better economic backgrounds. As seen in Table 2.5, non-disadvantaged black students are less likely to offend than their white counterparts. However, test scores of disadvantaged white students are higher than those of disadvantaged blacks students, meaning that the test score gap widens the residual gap in offending rate. When only considering students from economically disadvantaged backgrounds (columns (1), (3), and (5)), test scores explain a substantial portion of the offending gap – about 70% of the gap in misdemeanor charges and about 40% of the gap in felony charges. Moreover, for this population, we see that differential returns to test scores by race now explain a greater portion of the gap. In fact, for felonies, the differences in returns to test scores account for a greater fraction of the gap than the differences in levels.

In the appendix, we also show the decomposition of the black-white gap in convictions (Table 2.18) – similar to our previous analyses, the conviction results look very similar to the results for the felony gap. We also try restricting the sample to men (Tables 2.19 and 2.20) and observe similar explanatory power for test scores. We also test the explanatory power of just one set of test scores, specifically grade 8 math scores 2.21), and find that just using these scores explain about the same fraction of the raw offending gap.

2.5 Discussion and Conclusion

Similar to Neal and Johnson's (1996) findings on how test scores explain differences in earnings for blacks and whites, we find that differences in average scores for black and white students explain a substantial fraction of the difference in early offending rates for blacks and whites. This "skill gap" explains large fractions of black-white gaps in many long-run outcomes. This gap seems particularly relevant for the 1980s stagnation in the closing of the black-white wage gap, given the steep increases in returns to skill in the labor market over this period (O'Neill, 1990; Bound and Freeman, 1992).

With the strong caveat that our results are not causal, our findings suggest the possibility that the skill gap's impact on labor market outcomes for blacks and whites extends to differences in crime rates by race. Crime rates rose considerably in the 1980s and 1990s, concurrently with the rise in returns to skill (Western and Pettit, 2002; Raphael, 2006). Some of this rise may be due to lowered returns to formal employment. If workers are choosing between legal employment and criminal activity, then lower skilled individuals will find crime more valuable, particularly if the returns to skill in the latter are rising (Freeman, 1999). In this case, differences in average achievement for young black and white students may differentially incentivize the average black and white students' criminal propensity.

There are two empirical facts that complicate the above story about the relationship between cognitive skill and the black-white crime gap. First, while test scores help explain the overall residual crime gap, they explain a lower fraction of the differences in felony offending rates for blacks and whites. Again, this pattern is exacerbated when looking at students from worse economic backgrounds, who we would expect to be at greater risk for patterns. More serious offenses are likely to be a more natural substitute for formal employment,[9] suggesting that there is more to this story.

Second, average differences in test scores are not the only source of test scores' predictive power of the crime gap. The returns to higher test scores are significantly higher for black students, which accounts for a substantial portion of the gap, particularly for more serious offenses. Again, this is not causal, but suggests the differential treatment of blacks and whites in regards to skill. This discrimination could appear in two ways: indirectly through alternatives to crime such as legal employment, or directly through law enforcement.

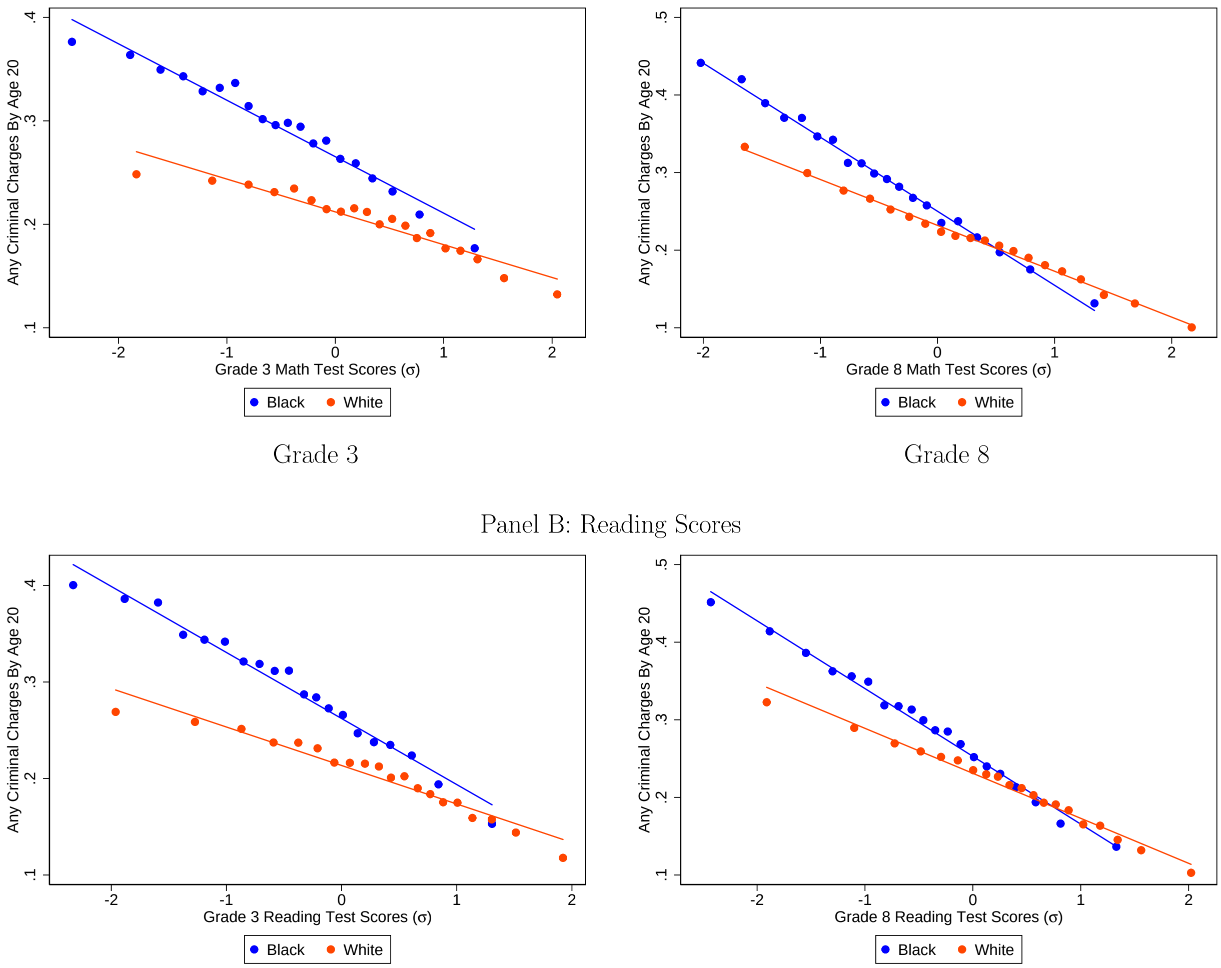
In the indirect mechanism, if there were higher returns to skill in the formal labor market for black workers, then we would also expect the crime test score gradient to also be steeper for black students (Freeman, 1999). There is weak evidence of differential returns to skill to wages by race for employed workers (Neal and Johnson, 1996) (although if anything, the returns are higher for black students (Neal, 2006), which would be consistent with this story). However, differential returns to skill could still be relevant on the extensive margin, where individuals decide to enter the labor market or engage in criminal activity. In the direct law enforcement mechanism, a more salient factor than test scores would likely need to be present, as cognitive skill is not directly observable.

As mentioned earlier, we emphasize that our estimates are not causal. The differences in levels of test scores themselves may reflect discrimination. The test scores we use could be racially biased (Jencks and Phillips, 1998) or black students may invest less in developing skills if they anticipate discrimination in the labor market (Lang and Manove, 2011). Likewise, the differences in returns to test scores may reflect differences in other factors related to test scores that affect black and white students differently. We plan to test our theories and these alternative explanations in future work.

2.6 Figures

[9] This seems especially true for North Carolina. Two major substitutes to legal income, drug sales and property crimes, are nearly always felonies. With the exception of smaller sales of certain drugs, the majority of drug sales are felony charges. Larceny worth over $1000 is always a felony.

Figure 2.1: Relationships Between Test Scores and Criminal Charge Rates by Race



Panel A: Math Scores

Grade 3

Grade 8

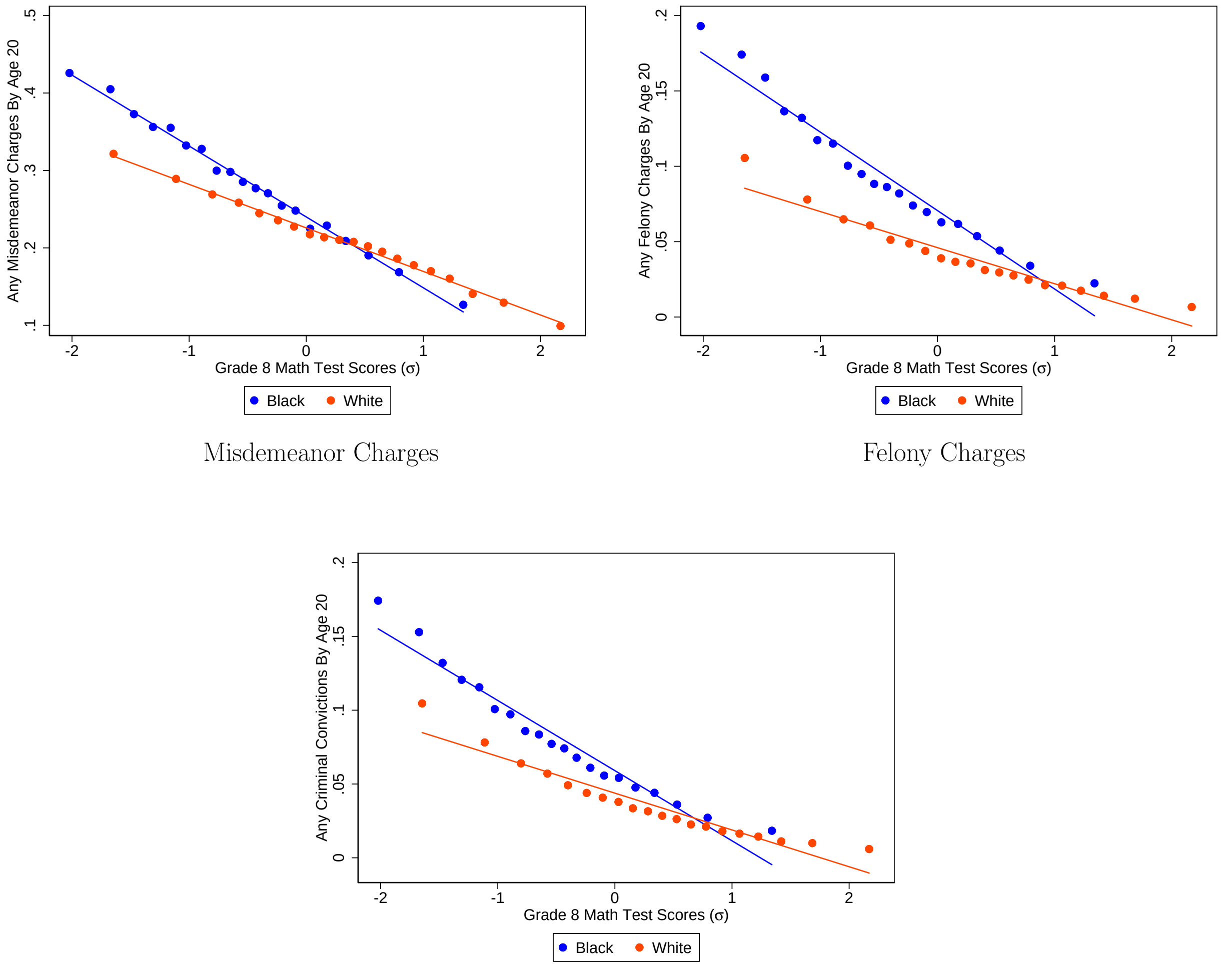
Panel B: Reading Scores

Grade 3

Grade 8

Notes: These figures display binned scatter plots of criminal charge rates by test score for each race. These figures were made by splitting black and white students into vingtiles of their race's respective test score distributions. Each point displays the average offending rate against the average test score within in each race-vingtile. The lines represent the bivariate relationship between charges and test scores for each race. We plot (i) criminal charge rates versus grade 3 math scores, (ii) criminal charge rates versus grade 8 math scores, (iii) criminal charge rates versus grade 3 reading scores, and (iv) criminal charge rates versus grade 8 reading scores. For our crime outcome, we focus on whether or not the student was charged with any offense between the ages of 16 and 20.

Figure 2.2: Relationships Between Test Scores and Crime by Crime Severity and Race



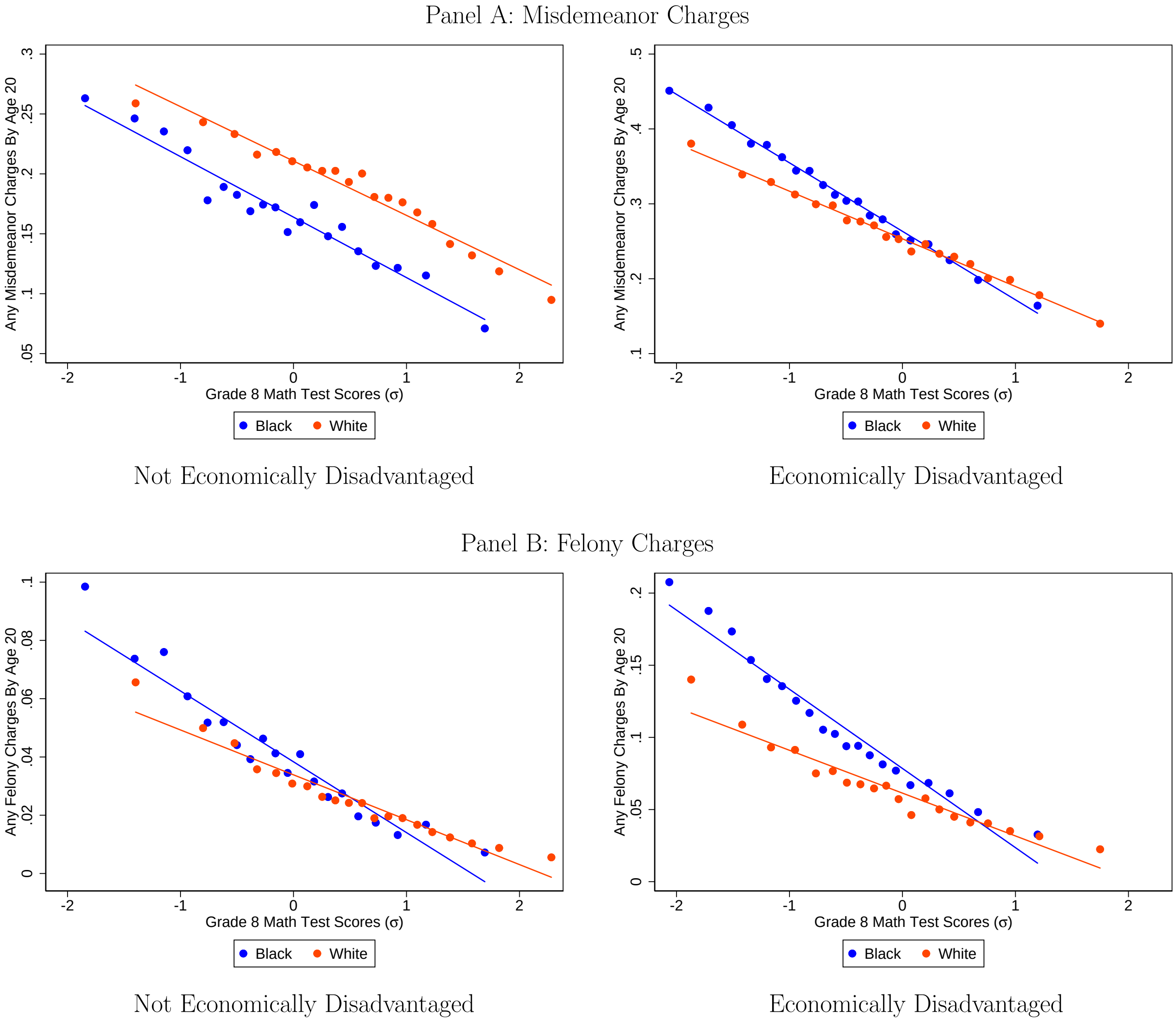
Misdemeanor Charges

Felony Charges

Convictions

Notes: These figures display binned scatter plots of crime rates of different crime severities against grade 8 math scores for each race. These figures were made by splitting black and white students into vingtiles of their race's respective test score distributions. Each point displays the average offending rate against the average test score within in each race-vingtile. The lines represent the bivariate relationship between charges and test scores for each race. The crime rates plotted here are (i) misdemeanor charge rates between ages 16 and 20, (ii) felony charge rates between ages 16 and 20, and (iii) conviction rates between ages 16 and 20.

Figure 2.3: Relationships Between Test Scores and Criminal Charge Rates by Race and Socio-Economic Status



Panel A: Misdemeanor Charges

Not Economically Disadvantaged

Economically Disadvantaged

Panel B: Felony Charges

Not Economically Disadvantaged

Economically Disadvantaged

Notes: These figures display binned scatter plots of crime rates of different crime severities against grade 8 math scores for each race. These plots were made separately for students from economically disadvantaged and non-disadvantaged backgrounds. These figures were made by splitting black and white students into vingtiles of their race's respective test score distributions. Each point displays the average offending rate against the average test score within in each race-vingtile. The lines represent the bivariate relationship between charges and test scores for each race.

2.7 Tables

Table 2.1: Summary Statistics

	Full Sample			All Offenders		
	(1) All	(2) Black	(3) White	(4) All	(5) Black	(6) White
Male	0.513	0.509	0.516	0.642	0.635	0.647
Black	0.337	1	0	0.418	1	0
White	0.663	0	1	0.582	0	1
Economically Disadvantaged	0.518	0.830	0.360	0.638	0.908	0.444
Special Ed	0.138	0.167	0.124	0.168	0.211	0.137
Parents Attended College	0.529	0.393	0.597	0.441	0.317	0.530
Graduated HS	0.737	0.692	0.760	0.577	0.524	0.616
Grade 3 Math	0.0258	-0.528	0.297	-0.190	-0.680	0.160
Grade 3 Read	0.0407	-0.452	0.282	-0.208	-0.653	0.107
Grade 8 Math	0.0995	-0.424	0.362	-0.199	-0.655	0.117
Grade 8 Read	0.0935	-0.428	0.354	-0.202	-0.674	0.124
Charged by Age 20	0.216	0.268	0.189			
Observations	803051	270326	532725	173241	72363	100878

Notes: This table displays descriptive statistics for black and white North Carolina public school students that were matched to criminal charge records. We selected our sample by restricting students who attended a public school between grades 3 and 8 and were born between 1989 and 1995. Column (1) represents the mean demographics of all students in this sample. Columns (2) and (3) provide the means of the black and white students in our sample, respectively. Column (4) reports the mean characteristics of anyone charged between ages 16 and 20, and Columns (5) and (6) provide the mean characteristics of black and white offenders, respectively. For economically disadvantaged status, special education enrollment, parents' college attendance status, and student's graduation status, we report the fraction of students that *ever* fell in said categories.

Table 2.2: The Residual Black-White Crime Gap in All Charges

	(1)	(2)	(3)	(4)	(5)	(6)
Black	0.0783***	0.0266***	0.0195***	0.0118***	0.0000805	0.0203***
	(0.00101)	(0.00105)	(0.00120)	(0.00127)	(0.00130)	(0.00128)
Test score controls		✓	✓	✓	✓	
Cohort FEs			✓	✓	✓	✓
School FEs			✓	✓	✓	✓
Census Tract FEs				✓	✓	✓
Family characteristics					✓	✓
White Offending Rate	0.189	0.189	0.189	0.189	0.189	0.189
Black Offending Rate	0.268	0.268	0.268	0.268	0.268	0.268
Test scores F-stat		1694.6	1442.4	1366.9	758.3	
R2	0.00810	0.0516	0.0687	0.0749	0.0943	0.0839
Total parameters	2	26	2233	5613	5625	5613
N	803051	803051	803051	803051	803051	803051

Notes: This table reports estimates of γ in Equation (2.1). The dependent variable is a binary variable indicating if the student was charged with any offense between the ages of 16 and 20. The test scores used here are all observed math and reading scores from grades 3 through 8. The schools used for the school fixed effects were the location that the most recent test was taken (usually grade 8). Census tract fixed effects were used for the tract that corresponded to the student's first observed bus stop. The family characteristics include the student's gender, the parents' highest level of education, and if the student ever received special education services or was reported to be economically disadvantaged. For each regressor, we included an additional binary variable control indicating if said value was missing, and then imputed means for the missing values. Robust standard errors are reported in parentheses.

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

Table 2.3: Residual Black-White Crime Gaps

	Any Charge			Misdemeanor Charge			Felony Charge		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Black	0.0203***	0.0000805	-0.00625***	0.0171***	-0.00215	-0.00822***	0.0228***	0.0150***	0.0112***
	(0.00128)	(0.00130)	(0.00133)	(0.00127)	(0.00129)	(0.00132)	(0.000706)	(0.000720)	(0.000735)
Test score controls		✓	✓		✓	✓		✓	✓
Test scores × Black			✓			✓			✓
Cohort FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
School FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Census Tract FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Family characteristics	✓	✓	✓	✓	✓	✓	✓	✓	✓
White Offending Rate	0.189	0.189	0.189	0.185	0.185	0.185	0.0358	0.0358	0.0358
Black Offending Rate	0.268	0.268	0.268	0.257	0.257	0.257	0.0899	0.0899	0.0899
Test scores F-stat		758.3	411.0		705.6	384.6		370.6	122.9
R2	0.0839	0.0943	0.0953	0.0794	0.0891	0.0901	0.0749	0.0800	0.0820
Total parameters	5612	5624	5636	5612	5624	5636	5612	5624	5636
N	803051	803051	803051	803051	803051	803051	803051	803051	803051

Notes: This table reports estimates of γ in Equation (2.1). The crime outcomes used in these regressions were binary variables indicating if the student was charged with any offense (Columns (1)-(3)), any misdemeanor offense (Columns (4)-(6)), and any felony offense (Columns (7)-(9)) between the ages of 16 and 20. The test score variables used here include all observed math and reading scores from grades 3 through 8. Columns (3), (6), and (9) also include interactions of these test scores with the black dummy as controls. The other controls used here are the same as those in the estimates from Table 2.2. Robust standard errors are reported in parentheses.

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

Table 2.4: Residual Black-White Crime Gaps for Economically Disadvantaged Students

	Any Charge			Misdemeanor Charge			Felony Charge		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Black	0.0395***	0.0200***	0.0119***	0.0358***	0.0173***	0.00954***	0.0291***	0.0204***	0.0151***
	(0.00182)	(0.00186)	(0.00193)	(0.00181)	(0.00184)	(0.00192)	(0.00114)	(0.00116)	(0.00121)
Test score controls		✓	✓		✓	✓		✓	✓
Test scores × Black			✓			✓			✓
Cohort FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
School FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Census Tract FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Family characteristics	✓	✓	✓	✓	✓	✓	✓	✓	✓
White Offending Rate	0.239	0.239	0.239	0.231	0.231	0.231	0.0602	0.0602	0.0602
Black Offending Rate	0.302	0.302	0.302	0.290	0.290	0.290	0.105	0.105	0.105
Test scores F-stat		405.8	159.9		379.3	150.9		209.0	52.62
R2	0.0891	0.100	0.101	0.0845	0.0951	0.0957	0.0856	0.0915	0.0926
Total parameters	4224	4236	4248	4224	4236	4248	4224	4236	4248
N	393709	393709	393709	393709	393709	393709	393709	393709	393709

Notes: This table reports estimates of γ in Equation (2.1), restricting our sample to economically disadvantaged students. The crime outcomes used in these regressions were binary variables indicating if the student was charged with any offense (Columns (1)-(3)), any misdemeanor offense (Columns (4)-(6)), and any felony offense (Columns (7)-(9)) between the ages of 16 and 20. The test score variables used here include all observed math and reading scores from grades 3 through 8. Columns (3), (6), and (9) also include interactions of these test scores with the black dummy as controls. The other controls used here are the same as those in the estimates from Table 2.2. Robust standard errors are reported in parentheses.

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

Table 2.5: Residual Black-White Crime Gaps for Economically Non-Disadvantaged Students

	Any Charge			Misdemeanor Charge			Felony Charge		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Black	-0.0163***	-0.0346***	-0.0347***	-0.0186***	-0.0360***	-0.0364***	0.00894***	0.00377***	0.00483***
	(0.00217)	(0.00219)	(0.00223)	(0.00215)	(0.00217)	(0.00222)	(0.000914)	(0.000926)	(0.000945)
Test score controls		✓	✓		✓	✓		✓	✓
Test scores × Black			✓			✓			✓
Cohort FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
School FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Census Tract FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Family characteristics	✓	✓	✓	✓	✓	✓	✓	✓	✓
White Offending Rate	0.168	0.168	0.168	0.165	0.165	0.165	0.0233	0.0233	0.0233
Black Offending Rate	0.150	0.150	0.150	0.144	0.144	0.144	0.0349	0.0349	0.0349
Test scores F-stat		325.1	296.4		305.0	279.5		145.9	119.5
R2	0.0538	0.0639	0.0639	0.0525	0.0620	0.0620	0.0325	0.0371	0.0373
Total parameters	4322	4334	4346	4322	4334	4346	4322	4334	4346
N	366932	366932	366932	366932	366932	366932	366932	366932	366932

Notes: This table reports estimates of γ in Equation (2.1), restricting our sample to non-disadvantaged students. The crime outcomes used in these regressions were binary variables indicating if the student was charged with any offense (Columns (1)-(3)), any misdemeanor offense (Columns (4)-(6)), and any felony offense (Columns (7)-(9)) between the ages of 16 and 20. The test score variables used here include all observed math and reading scores from grades 3 through 8. Columns (3), (6), and (9) also include interactions of these test scores with the black dummy as controls. The other controls used here are the same as those in the estimates from Table 2.2. Robust standard errors are reported in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.6: Blinder-Oaxaca Decomposition of the Black-White Crime Gap at Age 20

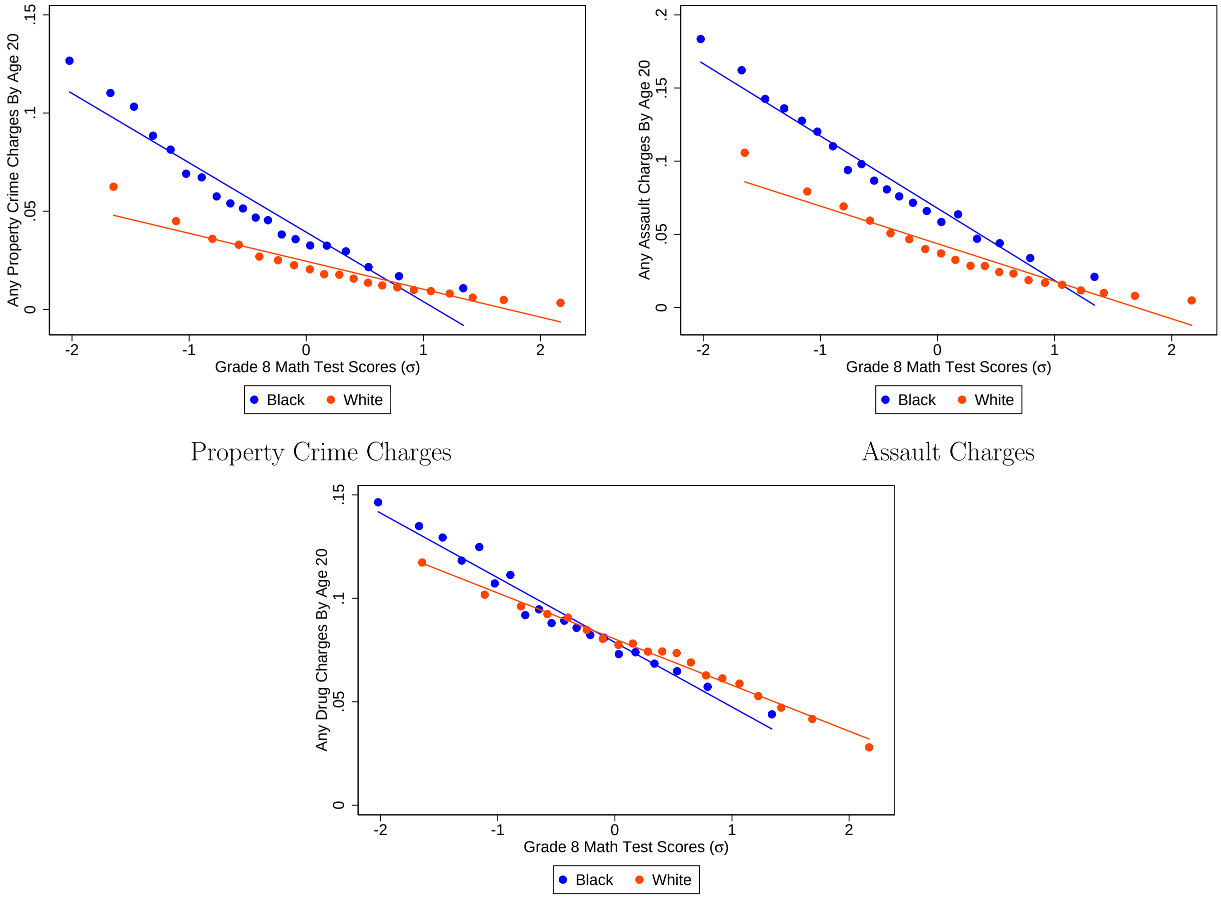
	Any Charge			Misdemeanor Charge			Felony Charge		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	All	Grade 8	Grade 3	All	Grade 8	Grade 3	All	Grade 8	Grade 3
Levels in Test Scores	0.0382***	0.0432***	0.0161***	0.0365***	0.0415***	0.0152***	0.0110***	0.0129***	0.0045***
	(0.0008)	(0.0007)	(0.0006)	(0.0008)	(0.0006)	(0.0006)	(0.0004)	(0.0003)	(0.0003)
	[46.01]	[54.45]	[17.75]	[48.08]	[57.10]	[18.07]	[20.12]	[23.43]	[7.54]
Returns in Test Scores	0.0141***	0.0125***	0.0105***	0.0136***	0.0120***	0.0103***	0.0081***	0.0106***	0.0068***
	(0.0008)	(0.0006)	(0.0007)	(0.0007)	(0.0006)	(0.0007)	(0.0004)	(0.0004)	(0.0004)
	[16.98]	[15.70]	[11.55]	[17.89]	[16.46]	[12.22]	[14.91]	[19.35]	[11.34]
Levels in Controls	0.0414***	0.0399***	0.0605***	0.0393***	0.0377***	0.0577***	0.0223***	0.0229***	0.0317***
	(0.0013)	(0.0011)	(0.0011)	(0.0013)	(0.0011)	(0.0011)	(0.0007)	(0.0006)	(0.0006)
	[49.84]	[50.27]	[66.66]	[51.71]	[51.81]	[68.64]	[40.81]	[41.71]	[53.25]
Residual	-0.0107***	-0.0161***	0.0037*	-0.0134***	-0.0184***	0.0009	0.0132***	0.0085***	0.0166***
	(0.0019)	(0.0016)	(0.0016)	(0.0019)	(0.0015)	(0.0016)	(0.0010)	(0.0009)	(0.0009)
	[-12.91]	[-20.33]	[4.12]	[-17.60]	[-25.26]	[1.09]	[24.13]	[15.53]	[27.88]
Cohort FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
School FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Census Tract FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Family Characteristics	✓	✓	✓	✓	✓	✓	✓	✓	✓
White Offending Rate	.2172	.2105	.2023	.2121	.2054	.1973	.0372	.0373	.0377
Black Offending Rate	.3003	.2899	.293	.288	.2781	.2813	.0918	.0922	.0973
N	483010	633552	614000	483010	633552	614000	483010	633552	614000

Notes: This table reports the estimates of each term of the Blinder-Oaxaca decomposition of the black-white offending gap, as depicted in Equation (2.3). Estimates were made from OLS estimates of the pooled model from Equation (2.2). Columns (1)-(3) decompose the gap in any offending by age 20, Columns (4)-(6) decompose the gap in any misdemeanor charges, and Columns (7)-(9) decompose the gap in felony charges. We do this decomposition using all math and reading scores from grades (3-8) (Columns (1), (4), and (7)), grade 8 tests only (Columns (2), (5), and (8)), and grades 3 only (Columns (3), (6), and (9)). Samples were restricted to individuals with observed test scores. Each row represents a different term in the breakdown. The first row represents the portion explained by average differences in test scores when using the white returns to test scores. The second row illustrates the difference in returns when using average test scores of black students test scores, the third row displays differences in S, and the fourth row displays the residual gap for the average student. The standard errors are in parentheses were calculated applying the delta method with robust standard error regression coefficients, assuming non-stochastic regressors (a la Oaxaca and Ransom (1998)). Each term's percent share of the raw black-white offending gap is reported below the standard errors in brackets.

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

2.8 Appendix: Additional Figures

Figure 2.4: Relationships Between Test Scores and Criminal Charge Rates by Race and Type of Criminal Offense



Property Crime Charges

Assault Charges

Drug Crime Charges

Notes: These figures display binned scatter plots of criminal charge rates for different crime types against grade 8 math scores for each race. These figures were made by splitting black and white students into vingtiles of their race's respective test score distributions. Each point displays the average offending rate against the average test score within in each race-vingtile. The lines represent the bivariate relationship between charges and test scores for each race. The property crime charge rates plotted here are (i) drug crime charge rates between ages 16 and 20, (ii) assault charge rates between ages 16 and 20, and (iii) drug crime rates between ages 16 and 20.

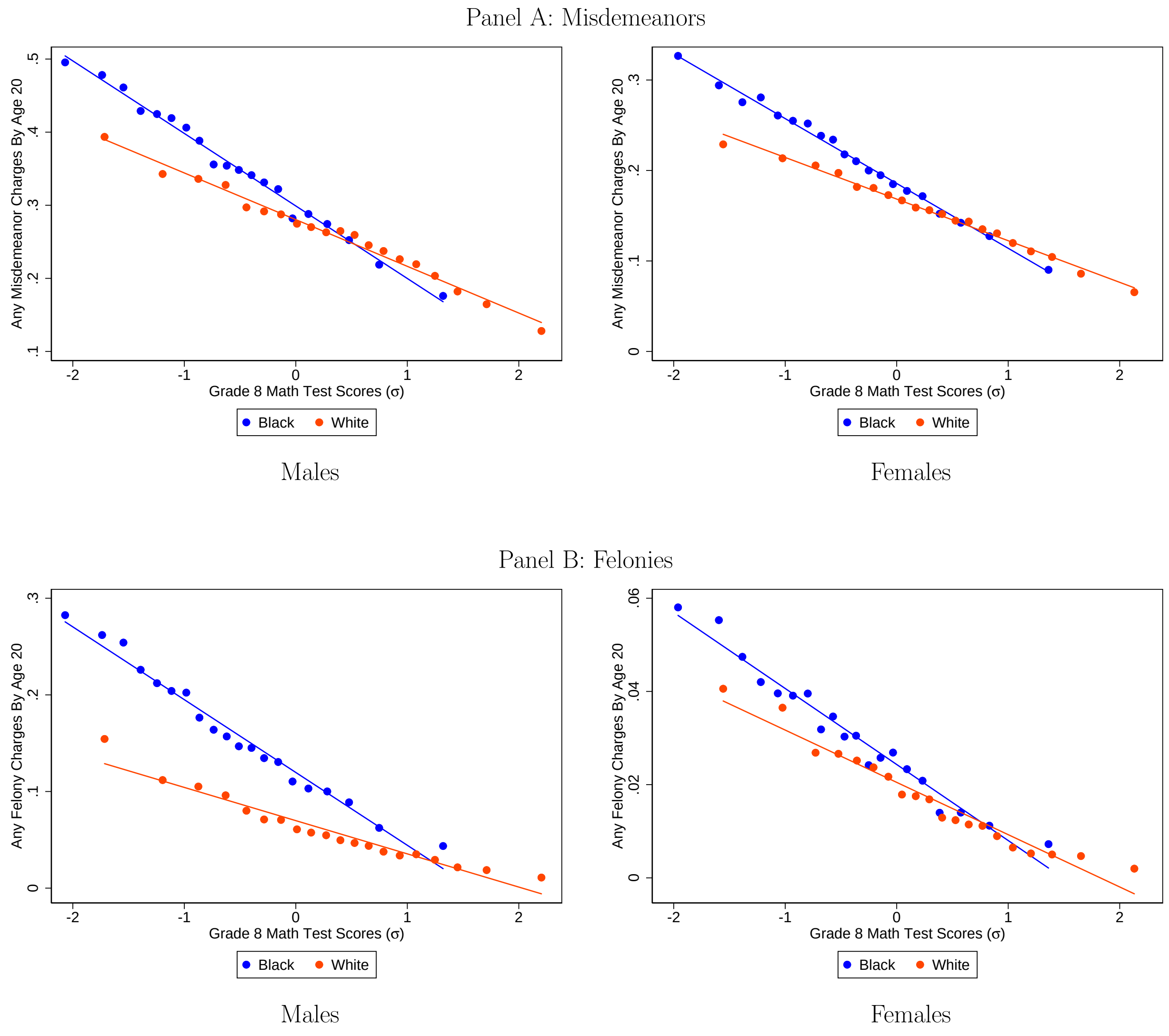
Table 2.7: Blinder-Oaxaca Decomposition of the Black-White Crime Gap by Economic Status

	Any Charge		Misdemeanor Charge		Felony Charge	
	(1)	(2)	(3)	(4)	(5)	(6)
	Low SES	High SES	Low SES	High SES	Low SES	High SES
Levels in Test Scores	0.0216***	0.0266***	0.0204***	0.0256***	0.0080***	0.0067***
	(0.0009)	(0.0007)	(0.0009)	(0.0007)	(0.0005)	(0.0003)
	[38.92]	[-143.24]	[40.39]	[-118.33]	[19.86]	[55.27]
Returns in Test Scores	0.0158***	0.0017*	0.0155***	0.0020**	0.0086***	-0.0013**
	(0.0012)	(0.0008)	(0.0012)	(0.0008)	(0.0007)	(0.0004)
	[28.55]	[-9.37]	[30.71]	[-9.46]	[21.30]	[-10.56]
Levels in Controls	0.0126***	0.0006	0.0121***	0.0005	0.0099***	0.0013
	(0.0016)	(0.0016)	(0.0016)	(0.0016)	(0.0010)	(0.0007)
	[22.79]	[-3.43]	[23.90]	[-2.54]	[24.52]	[10.26]
Residual	0.0054*	-0.0477***	0.0025	-0.0498***	0.0139***	0.0055***
	(0.0025)	(0.0032)	(0.0025)	(0.0032)	(0.0014)	(0.0016)
	[9.72]	[256.18]	[5.01]	[230.46]	[34.37]	[45.24]
Cohort FEs	✓	✓	✓	✓	✓	✓
School FEs	✓	✓	✓	✓	✓	✓
Census Tract FEs	✓	✓	✓	✓	✓	✓
Family Characteristics	✓	✓	✓	✓	✓	✓
White Offending Rate	.2685	.1901	.2602	.1866	.0615	.0243
Black Offending Rate	.324	.1715	.3107	.165	.1019	.0365
N	245531	237463	245531	237463	245531	237463

Notes: This table reports the estimates of each term of the Blinder-Oaxaca decomposition depicted in Equation (2.3) separately by students' socio-economic status. Estimates were made from OLS estimates of the pooled model from Equation (2.2). Columns (1), (3), and (5) restrict our sample to economically disadvantaged students, and Columns (2), (4), and (6) only include economically non-disadvantaged students. Columns (1)-(2) decompose the gap in any offending by age 20, Columns (3)-(4) decompose the gap in any misdemeanor charges, and Columns (5)-(6) decompose the gap in felony charges. We do this decomposition using all math and reading scores from grades (3-8). Samples were restricted to individuals with observed test scores. Each row represents a different term in the breakdown. The first row represents the portion explained by average differences in test scores when using the white returns to test scores. The second row illustrates the difference in returns when using average test scores of black students test scores, the third row displays differences in S, and the fourth row displays the residual gap for the average student. The standard errors are in parentheses were calculated applying the delta method with robust standard error regression coefficients, assuming non-stochastic regressors (a la Oaxaca and Ransom (1998)). Each term's percent share of the raw black-white offending gap is reported below the standard errors in brackets.

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

Figure 2.5: Relationships Between Test Scores and Criminal Charge Rates by Race and Gender



Panel A: Misdemeanors

Males

Females

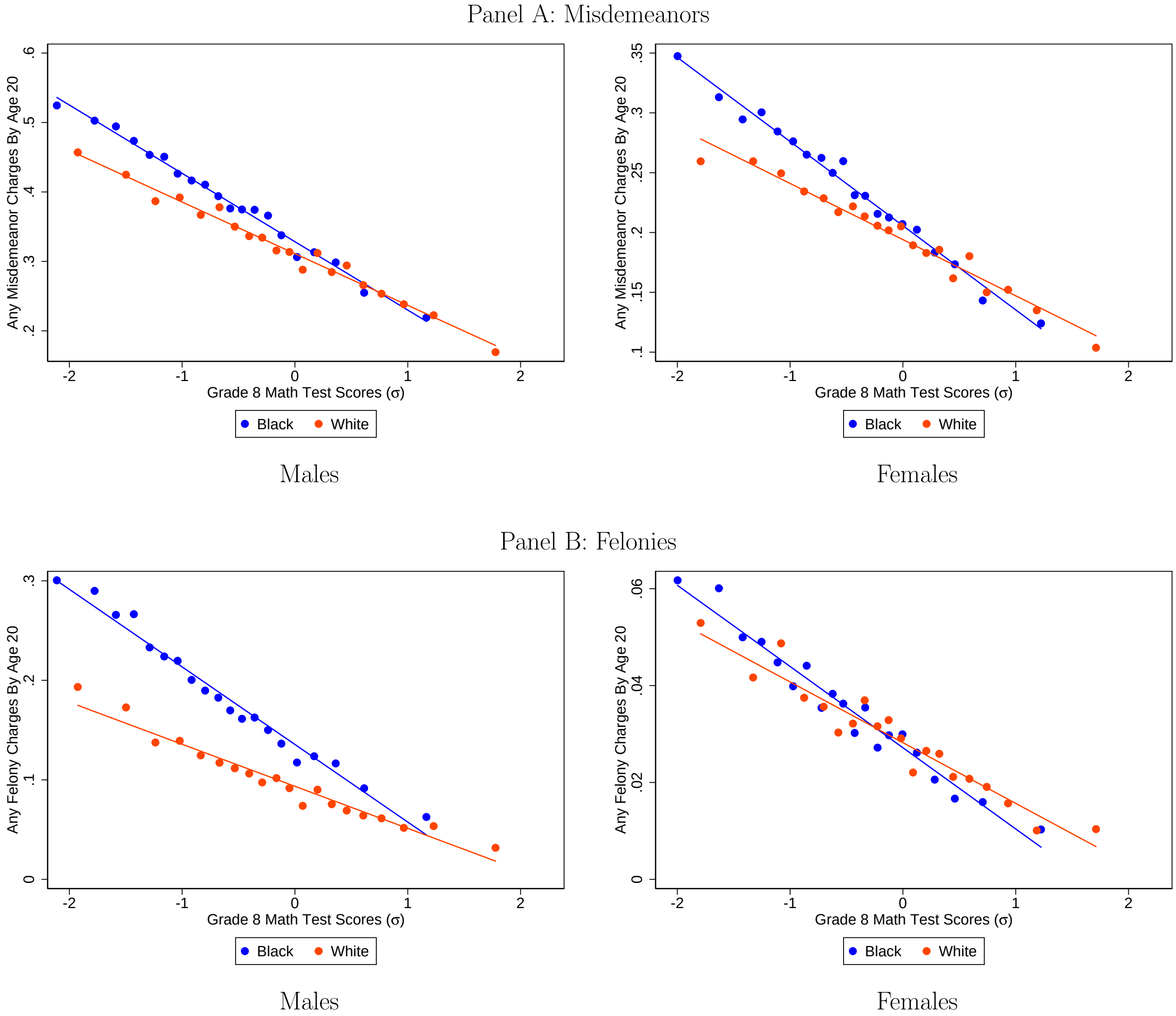
Panel B: Felonies

Males

Females

Notes: These figures display binned scatter plots of crime rates of different crime severities against grade 8 math scores for each race. These plots were made separately for male and female students. These figures were made by splitting black and white students into vingtiles of their race's respective test score distributions. Each point displays the average offending rate against the average test score within in each race-vingtile. The lines represent the bivariate relationship between charges and test scores for each race.

Figure 2.6: Relationships Between Test Scores and Criminal Charge Rates by Race and Gender for Economically Disadvantaged Students



Panel A: Misdemeanors

Males

Females

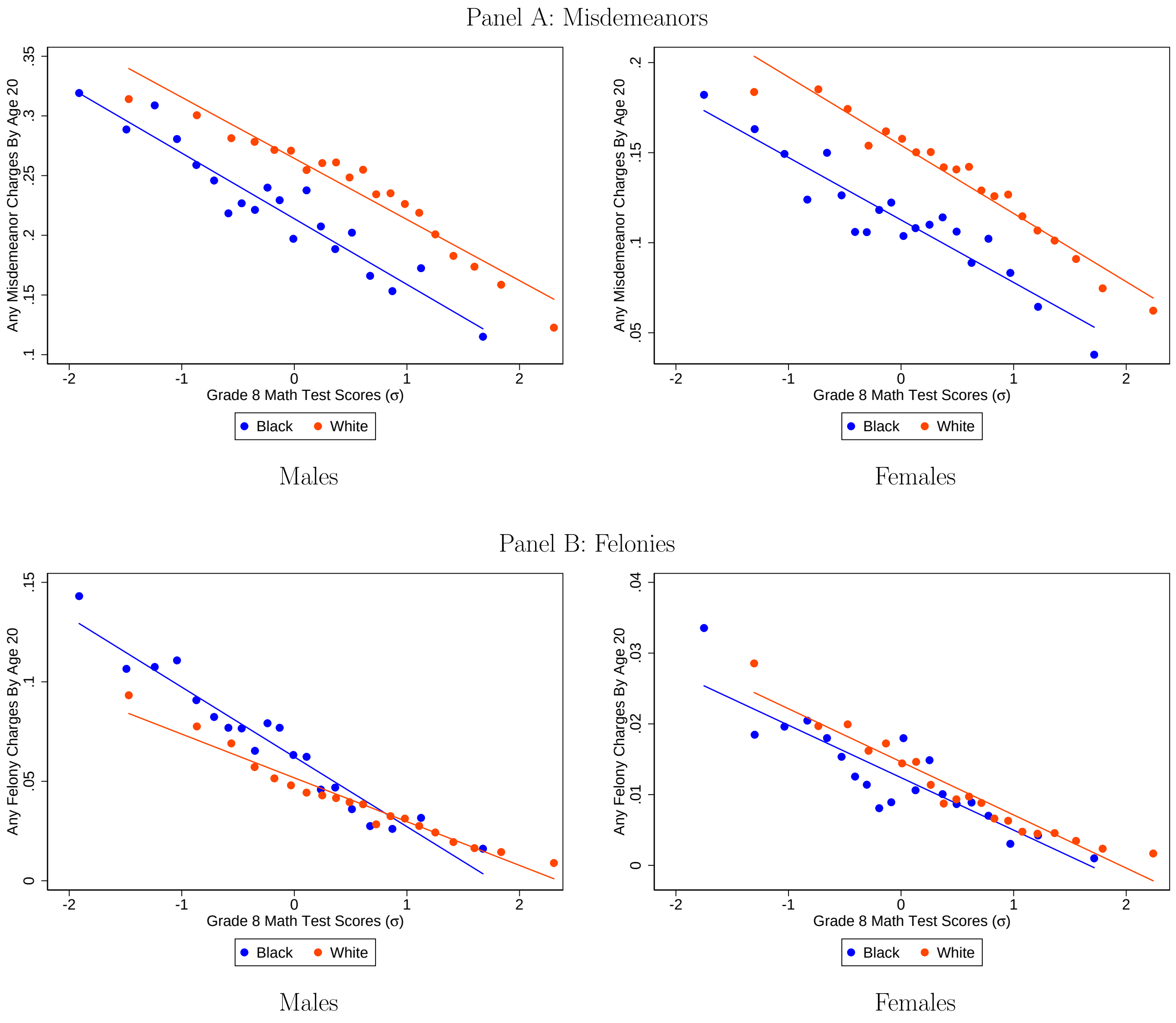
Panel B: Felonies

Males

Females

Notes: These figures display binned scatter plots of crime rates of different crime severities against grade 8 math scores for each race. These plots were made separately for male and female students, and we restricted the samples used to generate these plots to students from economically disadvantaged backgrounds. These figures were made by splitting black and white students into vingtiles of their race's respective test score distributions. Each point displays the average offending rate against the average test score within in each race-vingtile. The lines represent the bivariate relationship between charges and test scores for each race.

Figure 2.7: Relationships Between Test Scores and Criminal Charge Rates by Race and Gender for Economically Non-Disadvantaged Students


Panel A: Misdemeanors

Males

Females

Panel B: Felonies

Males

Females

Notes: These figures display binned scatter plots of crime rates of different crime severities against grade 8 math scores for each race. These plots were made separately for male and female students, and we restricted the samples used to generate these plots to students from economically non-disadvantaged backgrounds. These figures were made by splitting black and white students into vingtiles of their race's respective test score distributions. Each point displays the average offending rate against the average test score within in each race-vingtile. The lines represent the bivariate relationship between charges and test scores for each race.

2.9 Appendix: Additional Tables

Table 2.8: Summary Statistics for Students With Observed Convictions

	Full Sample			All Convicts		
	(1) All	(2) Black	(3) White	(4) All	(5) Black	(6) White
Male	0.514	0.509	0.517	0.804	0.810	0.798
Black	0.322	1	0	0.529	1	0
White	0.678	0	1	0.471	0	1
Economically Disadvantaged	0.485	0.803	0.328	0.765	0.921	0.584
Special Ed	0.135	0.165	0.121	0.264	0.295	0.229
Parents Attended College	0.503	0.370	0.566	0.287	0.233	0.348
Graduated HS	0.724	0.678	0.748	0.274	0.263	0.287
Grade 3 Math	0.0272	-0.528	0.286	-0.499	-0.816	-0.134
Grade 3 Read	0.0357	-0.459	0.266	-0.560	-0.844	-0.234
Grade 8 Math	0.0495	-0.494	0.306	-0.620	-0.902	-0.317
Grade 8 Read	0.0507	-0.475	0.299	-0.652	-0.960	-0.321
Convicted by Age 20	0.0442	0.0725	0.0307			
Observations	1965359	633648	1331711	86772	45918	40854

Notes: This table displays descriptive statistics for black and white North Carolina public school students that were matched to conviction records that required mandatory supervision as part of the sentence. We selected our sample by restricting students who attended a public school between grades 3 and 8 after the 1992-1993 school year and were born no later than 1995. Column (1) represents the mean demographics of all students in this sample. Columns (2) and (3) provide the means of the black and white students in our sample, respectively. Column (4) reports the mean characteristics of anyone convicted between ages 16 and 20, and Columns (5) and (6) provide the mean characteristics of black and white offenders, respectively. The demographic variables are the same as the variables described in Table 2.1.

Table 2.9: The Residual Black-White Crime Gap in Charges by Age 20 Using Different Controls and Only Grade 8 Test Scores

	(1)	(2)	(3)	(4)	(5)	(6)
Black	0.0795***	0.0167***	0.00945***	0.00106	-0.0113***	0.0169***
	(0.00117)	(0.00125)	(0.00142)	(0.00152)	(0.00156)	(0.00153)
Math test score		-0.0425***	-0.0459***	-0.0457***	-0.0443***	
		(0.000825)	(0.000850)	(0.000854)	(0.000864)	
Reading test score		-0.0376***	-0.0338***	-0.0334***	-0.0224***	
		(0.000833)	(0.000831)	(0.000834)	(0.000844)	
Cohort FEs			✓	✓	✓	✓
School FEs			✓	✓	✓	✓
Census Tract FEs				✓	✓	✓
Family characteristics					✓	✓
White Offending Rate	0.210	0.210	0.210	0.210	0.210	0.210
Black Offending Rate	0.290	0.290	0.290	0.290	0.290	0.290
Test scores F-stat		8449.3	7144.7	6831.3	4004.5	
R2	0.00776	0.0327	0.0519	0.0581	0.0799	0.0682
Total parameters	2	4	877	4195	4207	4205
N	633552	633552	633552	633552	633552	633552

Notes: This table reports estimates of γ in Equation (2.1). The dependent variable is a binary variable indicating if the student was charged with any offense between the ages of 16 and 20. The test scores used here are all observed math and reading scores from grade 8 only. The other controls used here are the same as those in the estimates from Table 2.2. Robust standard errors are reported in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2.10: Magnitude of the Residual Black-White Gap in the NLSY97 for Arrests by Age 20 Using Different Controls

	(1)	(2)	(3)	(4)	(5)	(6)
Black	0.0500***	0.00827	0.00856	-0.0141	0.00669	-0.0269
	(0.0112)	(0.0121)	(0.0121)	(0.0125)	(0.0121)	(0.0139)
Test score controls		✓	✓	✓		✓
Test scores × Black						✓
F-stat for Test Scores				23.94		13.32
Cohort FEs			✓	✓	✓	✓
School & City Controls				✓	✓	✓
Family characteristics				✓	✓	✓
White Offending Rate	0.187	0.187	0.187	0.187	0.187	0.187
Black Offending Rate	0.237	0.237	0.237	0.237	0.237	0.237
R2	0.00233	0.0183	0.0189	0.0746	0.0680	0.0749
Total parameters	2	4	8	20	18	22
N	7000	7000	7000	7000	7000	7000

Notes: This table reports estimates of γ in Equation (2.1) when using the NLSY97. The dependent variable is a binary variable indicating if the student was ever arrested between the ages of 16 and 20. The test scores used here are weighted combinations of the ASVAB tests designed to mimic the AFQT, standardized to have mean zero and standard deviation 1. Column (6) also includes an interaction of these test scores with the black dummy variable as an additional regressor. The school and city controls include a dummy variable for if the student attended a high school in an urban setting during the first interview, and whether or not there were gangs in said school. Family characteristics include the student's gender, a quadratic in the poverty to income ratio, and the mother's and father's highest level of education. For each regressor, we included an additional binary variable control indicating if said value was missing, and then imputed means for the missing values. Robust standard errors are reported in parentheses.

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

Table 2.11: Magnitude of the Residual Black-White Gap in the NLSY97 for Incarceration by Age 20 Using Different Controls

	(1)	(2)	(3)	(4)	(5)	(6)
Black	0.00973***	0.00738*	0.00747*	0.00575	0.00672*	0.00270
	(0.00282)	(0.00309)	(0.00309)	(0.00323)	(0.00299)	(0.00390)
Test score controls		✓	✓	✓		✓
Test scores × Black						✓
F-stat for Test Scores				4.844		2.671
Cohort FEs			✓	✓	✓	✓
School & City Controls				✓	✓	✓
Family characteristics				✓	✓	✓
White Offending Rate	0.00413	0.00413	0.00413	0.00413	0.00413	0.00413
Black Offending Rate	0.0139	0.0139	0.0139	0.0139	0.0139	0.0139
R2	0.00239	0.00544	0.00685	0.0120	0.00986	0.0124
Total parameters	2	4	8	20	18	22
N	7000	7000	7000	7000	7000	7000

Notes: This table reports estimates of γ in Equation (2.1). The dependent variable is a binary variable indicating if the student was ever incarcerated between the ages of 16 and 20. The test scores used here are weighted combinations of the ASVAB tests designed to mimic the AFQT, standardized to have mean zero and standard deviation 1. Column (6) also includes an interaction of these test scores with the black dummy variable as an additional regressor. The school and city controls include a dummy variable for if the student attended a high school in an urban setting during the first interview, and whether or not there were gangs in said school. Family characteristics include the student's gender, a quadratic in the poverty to income ratio, and the mother's and father's highest level of education. For each regressor, we included an additional binary variable control indicating if said value was missing, and then imputed means for the missing values. Robust standard errors are reported in parentheses.

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

Table 2.12: The Residual Black-White Crime Gap in Charges by Age 20 When Controlling for Educational Attainment

	(1)	(2)	(3)	(4)	(5)	(6)
Black	0.0783***	0.0266***	0.0266***	0.0191***	0.0118***	0.0234***
	(0.00101)	(0.00105)	(0.00103)	(0.00125)	(0.00128)	(0.00126)
Test score controls		✓	✓	✓	✓	
Years of School			✓	✓	✓	✓
Cohort FEs				✓	✓	✓
School FEs				✓	✓	✓
Census Tract FEs				✓	✓	✓
Family characteristics					✓	✓
White Offending Rate	0.189	0.189	0.189	0.189	0.189	0.189
Black Offending Rate	0.268	0.268	0.268	0.268	0.268	0.268
Test scores F-stat		1694.6	431.0	399.0	263.3	
R2	0.00810	0.0516	0.0953	0.112	0.125	0.122
Total parameters	2	26	75	5629	5641	5629
N	803051	803051	803051	803051	803051	803051

Notes: This table reports estimates of γ in Equation (2.1). The dependent variable is a binary variable indicating if the student was charged with any offense between the ages of 16 and 20. The test scores used here are all observed math and reading scores from grades 3 through 8. We also include controls for educational attainment, including the maximum grade we observe the student attending in the North Carolina public school system, whether or not the student ever repeated a grade in elementary, middle, or high school, and whether or not the student graduated. The other controls used here are the same as those in the estimates from Table 2.2. Robust standard errors are reported in parentheses.

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

Table 2.13: Magnitude of the Residual Black-White Crime Gap at Age 20 By Crime Type and Types of Test Score Controls

	Drug Crime Charge			Assault Charge			Property Crime Charge		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Black	-0.000383	-0.00618***	-0.00698***	0.0230***	0.0153***	0.0116***	0.0148***	0.00989***	0.00693***
	(0.000817)	(0.000834)	(0.000852)	(0.000692)	(0.000705)	(0.000720)	(0.000543)	(0.000554)	(0.000565)
Test score controls		✓	✓		✓	✓		✓	✓
Test scores × Black			✓			✓			✓
Cohort FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
School FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Census Tract FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Family characteristics	✓	✓	✓	✓	✓	✓	✓	✓	✓
White Offending Rate	0.0653	0.0653	0.0653	0.0329	0.0329	0.0329	0.0188	0.0188	0.0188
Black Offending Rate	0.0857	0.0857	0.0857	0.0859	0.0859	0.0859	0.0539	0.0539	0.0539
Test scores F-stat		224.1	147.8		344.8	125.4		229.9	64.41
R2	0.0557	0.0588	0.0592	0.0602	0.0650	0.0664	0.0592	0.0625	0.0643
Total parameters	5612	5624	5636	5612	5624	5636	5612	5624	5636
N	803051	803051	803051	803051	803051	803051	803051	803051	803051

Notes: This table reports estimates of γ in Equation (2.1). The crime outcomes used in these regressions were binary variables indicating if the student was charged with any drug offense (Columns (1)-(3)), any assaults (Columns (4)-(6)), and any property offenses such as larceny (Columns (7)-(9)) between the ages of 16 and 20. The test score variables used here include all observed math and reading scores from grades 3 through 8. Columns (3), (6), and (9) also include interactions of these test scores with the black dummy as controls. The other controls used here are the same as those in the estimates from Table 2.2. Robust standard errors are reported in parentheses.

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

Table 2.14: Magnitude of the Residual Black-White Crime Gap for Men at Age 20 By Charge Severity and Types of Test Score Controls

	Any Charge			Misdemeanor Charge			Felony Charge		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Black	0.0306***	0.00453*	-0.00566**	0.0250***	0.000276	-0.00924***	0.0441***	0.0310***	0.0210***
	(0.00193)	(0.00197)	(0.00203)	(0.00192)	(0.00195)	(0.00201)	(0.00122)	(0.00125)	(0.00129)
Test score controls		✓	✓		✓	✓		✓	✓
Test scores × Black			✓			✓			✓
Cohort FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
School FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Census Tract FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Family characteristics	✓	✓	✓	✓	✓	✓	✓	✓	✓
White Offending Rate	0.238	0.238	0.238	0.231	0.231	0.231	0.0551	0.0551	0.0551
Black Offending Rate	0.334	0.334	0.334	0.318	0.318	0.318	0.149	0.149	0.149
Test scores F-stat		440.6	248.4		404.5	230.2		276.8	92.88
R2	0.0860	0.0978	0.0988	0.0813	0.0922	0.0931	0.0854	0.0928	0.0955
Total parameters	5251	5263	5275	5251	5263	5275	5251	5263	5275
N	412221	412221	412221	412221	412221	412221	412221	412221	412221

Notes: This table reports estimates of γ in Equation (2.1) when restricting our sample to men. The crime outcomes used in these regressions were binary variables indicating if the student was charged with any offense (Columns (1)-(3)), any misdemeanor offense (Columns (4)-(6)), and any felony offense (Columns (7)-(9)) between the ages of 16 and 20. The test score variables used here include all observed math and reading scores from grades 3 through 8. Columns (3), (6), and (9) also include interactions of these test scores with the black dummy as controls. The other controls used here are the same as those in the estimates from Table 2.2. Robust standard errors are reported in parentheses.

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

Table 2.15: Magnitude of the Residual Black-White Crime Gap for Women at Age 20 By Charge Severity and Types of Test Score Controls

	Any Charge			Misdemeanor Charge			Felony Charge		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Black	0.00978***	-0.00478**	-0.00731***	0.00916***	-0.00487**	-0.00751***	0.000859	-0.00208**	-0.00193**
	(0.00167)	(0.00169)	(0.00173)	(0.00165)	(0.00168)	(0.00172)	(0.000645)	(0.000656)	(0.000672)
Test score controls		✓	✓		✓	✓		✓	✓
Test scores × Black			✓			✓			✓
Cohort FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
School FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Census Tract FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Family characteristics	✓	✓	✓	✓	✓	✓	✓	✓	✓
White Offending Rate	0.138	0.138	0.138	0.135	0.135	0.135	0.0152	0.0152	0.0152
Black Offending Rate	0.199	0.199	0.199	0.193	0.193	0.193	0.0285	0.0285	0.0285
Test scores F-stat		302.3	160.7		287.2	151.8		88.15	45.58
R2	0.0571	0.0659	0.0669	0.0556	0.0640	0.0649	0.0255	0.0282	0.0285
Total parameters	5103	5115	5127	5103	5115	5127	5103	5115	5127
N	390830	390830	390830	390830	390830	390830	390830	390830	390830

Notes: This table reports estimates of γ in Equation (2.1) when restricting our sample to women. The crime outcomes used in these regressions were binary variables indicating if the student was charged with any offense (Columns (1)-(3)), any misdemeanor offense (Columns (4)-(6)), and any felony offense (Columns (7)-(9)) between the ages of 16 and 20. The test score variables used here include all observed math and reading scores from grades 3 through 8. Columns (3), (6), and (9) also include interactions of these test scores with the black dummy as controls. The other controls used here are the same as those in the estimates from Table 2.2. Robust standard errors are reported in parentheses.

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

Table 2.16: Blinder-Oaxaca Decomposition of the Black-White Crime Gap in Misdemeanors at Age 20 Using Test Scores in Different Years

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	All Tests	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Levels in Test Scores	0.0365***	0.0152***	0.0201***	0.0222***	0.0293***	0.0327***	0.0415***
	(0.0008)	(0.0006)	(0.0006)	(0.0006)	(0.0007)	(0.0006)	(0.0006)
	[48.08]	[18.07]	[24.28]	[27.29]	[37.28]	[43.41]	[57.10]
Returns in Test Scores	0.0136***	0.0103***	0.0120***	0.0115***	0.0132***	0.0126***	0.0120***
	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0006)	(0.0006)
	[17.89]	[12.22]	[14.52]	[14.16]	[16.83]	[16.75]	[16.46]
Levels in Controls	0.0393***	0.0577***	0.0544***	0.0515***	0.0454***	0.0430***	0.0377***
	(0.0013)	(0.0011)	(0.0011)	(0.0011)	(0.0011)	(0.0011)	(0.0011)
	[51.71]	[68.64]	[65.75]	[63.32]	[57.70]	[57.01]	[51.81]
Residual	-0.0134***	0.0009	-0.0038*	-0.0039*	-0.0093***	-0.0129***	-0.0184***
	(0.0019)	(0.0016)	(0.0016)	(0.0016)	(0.0016)	(0.0016)	(0.0015)
	[-17.60]	[1.09]	[-4.59]	[-4.76]	[-11.84]	[-17.15]	[-25.26]
Cohort FEs	✓	✓	✓	✓	✓	✓	✓
School FEs	✓	✓	✓	✓	✓	✓	✓
Census Tract FEs	✓	✓	✓	✓	✓	✓	✓
Family Characteristics	✓	✓	✓	✓	✓	✓	✓
White Offending Rate	.2121	.1973	.2007	.2035	.2058	.2065	.2054
Black Offending Rate	.288	.2813	.2835	.2848	.2845	.2819	.2781
N	483010	614000	616934	621492	625518	630565	633552

Notes: This table reports the estimates of each term of the Blinder-Oaxaca decomposition of the black-white misdemeanor offending gap, as depicted in Equation (2.3), using test scores for different grades. Estimates were made from OLS estimates of the pooled model from Equation (2.2). We do this decomposition using math and reading scores from all grades (Column (1)) and each individual grade separately (Columns (2)-(7)). Samples were restricted to individuals with observed test scores. Each row represents a different term in the breakdown. The first row represents the portion explained by average differences in test scores when using the white returns to test scores. The second row illustrates the difference in returns when using average test scores of black students test scores, the third row displays differences in S, and the fourth row displays the residual gap for the average student. The standard errors are in parentheses were calculated applying the delta method with robust standard error regression coefficients, assuming non-stochastic regressors (a la Oaxaca and Ransom (1998)). Each term's percent share of the raw black-white offending gap is reported below the standard errors in brackets.

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

Table 2.17: Blinder-Oaxaca Decomposition of the Black-White Crime Gap in Felonies at Age 20 Using Test Scores in Different Years

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	All Tests	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Levels in Test Scores	0.0110***	0.0045***	0.0061***	0.0067***	0.0096***	0.0098***	0.0129***
	(0.0004)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
	[20.12]	[7.54]	[10.35]	[11.37]	[16.30]	[17.28]	[23.43]
Returns in Test Scores	0.0081***	0.0068***	0.0086***	0.0087***	0.0106***	0.0103***	0.0106***
	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
	[14.91]	[11.34]	[14.50]	[14.67]	[18.10]	[18.08]	[19.35]
Levels in Controls	0.0223***	0.0317***	0.0307***	0.0300***	0.0272***	0.0260***	0.0229***
	(0.0007)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)
	[40.81]	[53.25]	[51.70]	[50.76]	[46.36]	[45.62]	[41.71]
Residual	0.0132***	0.0166***	0.0139***	0.0137***	0.0114***	0.0108***	0.0085***
	(0.0010)	(0.0009)	(0.0009)	(0.0009)	(0.0009)	(0.0009)	(0.0009)
	[24.13]	[27.88]	[23.41]	[23.21]	[19.37]	[18.97]	[15.53]
Cohort FEs	✓	✓	✓	✓	✓	✓	✓
School FEs	✓	✓	✓	✓	✓	✓	✓
Census Tract FEs	✓	✓	✓	✓	✓	✓	✓
Family Characteristics	✓	✓	✓	✓	✓	✓	✓
White Offending Rate	.0372	.0377	.0384	.0387	.0389	.0383	.0373
Black Offending Rate	.0918	.0973	.0978	.0978	.0975	.0953	.0922
N	483010	614000	616934	621492	625518	630565	633552

Notes: This table reports the estimates of each term of the Blinder-Oaxaca decomposition of the black-white felony offending gap, as depicted in Equation (2.3), using test scores for different grades. Estimates were made from OLS estimates of the pooled model from Equation (2.2). We do this decomposition using math and reading scores from all grades (Column (1)) and each individual grade separately (Columns (2)-(7)). Samples were restricted to individuals with observed test scores. Each row represents a different term in the breakdown. The first row represents the portion explained by average differences in test scores when using the white returns to test scores. The second row illustrates the difference in returns when using average test scores of black students test scores, the third row displays differences in S, and the fourth row displays the residual gap for the average student. The standard errors are in parentheses were calculated applying the delta method with robust standard error regression coefficients, assuming non-stochastic regressors (a la Oaxaca and Ransom (1998)). Each term's percent share of the raw black-white offending gap is reported below the standard errors in brackets.

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

Table 2.18: Blinder-Oaxaca Decomposition of the Black-White Crime Gap in Convictions at Age 20 Using Test Scores in Different Years

	(1) All Tests	(2) Grade 3	(3) Grade 4	(4) Grade 5	(5) Grade 6	(6) Grade 7	(7) Grade 8
Levels in Test Scores	0.0122***	0.0042***	0.0053***	0.0063***	0.0092***	0.0107***	0.0136***
	(0.0003)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
	[25.65]	[8.65]	[11.59]	[13.52]	[19.68]	[23.51]	[29.66]
Returns in Test Scores	0.0090***	0.0067***	0.0075***	0.0082***	0.0097***	0.0094***	0.0102***
	(0.0004)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
	[18.85]	[13.68]	[16.27]	[17.50]	[20.80]	[20.52]	[22.24]
Levels in Controls	0.0194***	0.0287***	0.0266***	0.0259***	0.0229***	0.0199***	0.0168***
	(0.0006)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)
	[40.57]	[58.89]	[57.98]	[55.17]	[49.31]	[43.60]	[36.47]
Residual	0.0071***	0.0092***	0.0065***	0.0064***	0.0047***	0.0056***	0.0053***
	(0.0008)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0005)	(0.0005)
	[14.99]	[18.88]	[14.21]	[13.74]	[10.21]	[12.35]	[11.45]
Cohort FEs	✓	✓	✓	✓	✓	✓	✓
School FEs	✓	✓	✓	✓	✓	✓	✓
Census Tract FEs	✓	✓	✓	✓	✓	✓	✓
Family Characteristics	✓	✓	✓	✓	✓	✓	✓
White Offending Rate	.0368	.0334	.0324	.0342	.0346	.0352	.036
Black Offending Rate	.0845	.0822	.0782	.0811	.0811	.0809	.082
N	683618	989218	1086117	1060401	1173656	1274112	1299412

Notes: This table reports the estimates of each term of the Blinder-Oaxaca decomposition of the black-white convictions offending gap, as depicted in Equation (2.3), using test scores for different grades. Estimates were made from OLS estimates of the pooled model from Equation (2.2). We do this decomposition using math and reading scores from all grades (Column (1)) and each individual grade separately (Columns (2)-(7)). Samples were restricted to individuals with observed test scores. Each row represents a different term in the breakdown. The first row represents the portion explained by average differences in test scores when using the white returns to test scores. The second row illustrates the difference in returns when using average test scores of black students test scores, the third row displays differences in S, and the fourth row displays the residual gap for the average student. The standard errors are in parentheses were calculated applying the delta method with robust standard error regression coefficients, assuming non-stochastic regressors (a la Oaxaca and Ransom (1998)). Each term's percent share of the raw black-white offending gap is reported below the standard errors in brackets.

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

Table 2.19: Blinder-Oaxaca Decomposition of Test Scores and the Black-White Crime Gap for Males at Age 20

	Any Charge			Misdemeanor Charge			Felony Charge		
	(1) All	(2) Grade 8	(3) Grade 3	(4) All	(5) Grade 8	(6) Grade 3	(7) All	(8) Grade 8	(9) Grade 3
Levels in Test Scores	0.0382***	0.0432***	0.0161***	0.0365***	0.0415***	0.0152***	0.0110***	0.0129***	0.0045***
	(0.0008)	(0.0007)	(0.0006)	(0.0008)	(0.0006)	(0.0006)	(0.0004)	(0.0003)	(0.0003)
	[46.01]	[54.45]	[17.75]	[48.08]	[57.10]	[18.07]	[20.12]	[23.43]	[7.54]
Returns in Test Scores	0.0141***	0.0125***	0.0105***	0.0136***	0.0120***	0.0103***	0.0081***	0.0106***	0.0068***
	(0.0008)	(0.0006)	(0.0007)	(0.0007)	(0.0006)	(0.0007)	(0.0004)	(0.0004)	(0.0004)
	[16.98]	[15.70]	[11.55]	[17.89]	[16.46]	[12.22]	[14.91]	[19.35]	[11.34]
Levels in Controls	0.0414***	0.0399***	0.0605***	0.0393***	0.0377***	0.0577***	0.0223***	0.0229***	0.0317***
	(0.0013)	(0.0011)	(0.0011)	(0.0013)	(0.0011)	(0.0011)	(0.0007)	(0.0006)	(0.0006)
	[49.84]	[50.27]	[66.66]	[51.71]	[51.81]	[68.64]	[40.81]	[41.71]	[53.25]
Residual	-0.0107***	-0.0161***	0.0037*	-0.0134***	-0.0184***	0.0009	0.0132***	0.0085***	0.0166***
	(0.0019)	(0.0016)	(0.0016)	(0.0019)	(0.0015)	(0.0016)	(0.0010)	(0.0009)	(0.0009)
	[-12.91]	[-20.33]	[4.12]	[-17.60]	[-25.26]	[1.09]	[24.13]	[15.53]	[27.88]
Cohort FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
School FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Census Tract FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Family Characteristics	✓	✓	✓	✓	✓	✓	✓	✓	✓
White Offending Rate	.2172	.2105	.2023	.2121	.2054	.1973	.0372	.0373	.0377
Black Offending Rate	.3003	.2899	.293	.288	.2781	.2813	.0918	.0922	.0973
N	483010	633552	614000	483010	633552	614000	483010	633552	614000

Notes: This table reports the estimates of each term of the Blinder-Oaxaca decomposition of the black-white offending gap for men, as depicted in Equation (2.3). Estimates were made from OLS estimates of the pooled model from Equation (2.2). Columns (1)-(3) decompose the gap in any offending by age 20, Columns (4)-(6) decompose the gap in any misdemeanor charges, and Columns (7)-(9) decompose the gap in felony charges. We do this decomposition using all math and reading scores from grades (3-8) (Columns (1), (4), and (7)), grade 8 tests only (Columns (2), (5), and (8)), and grades 3 only (Columns (3), (6), and (9)). Samples were restricted to individuals with observed test scores. Each row represents a different term in the breakdown. The first row represents the portion explained by average differences in test scores when using the white returns to test scores. The second row illustrates the difference in returns when using average test scores of black students test scores, the third row displays differences in S, and the fourth row displays the residual gap for the average student. The standard errors are in parentheses were calculated applying the delta method with robust standard error regression coefficients, assuming non-stochastic regressors (a la Oaxaca and Ransom (1998)). Each term's percent share of the raw black-white offending gap is reported below the standard errors in brackets.

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

Table 2.20: Blinder-Oaxaca Decomposition of Test Scores and the Black-White Crime Gap at Age 20 By Crime Type and Economic Status for Male Students

	Any Charge		Misdemeanor Charge		Felony Charge	
	(1)	(2)	(3)	(4)	(5)	(6)
	Low SES	High SES	Low SES	High SES	Low SES	High SES
Levels in Test Scores	0.0296***	0.0363***	0.0279***	0.0348***	0.0126***	0.0112***
	(0.0015)	(0.0012)	(0.0015)	(0.0012)	(0.0010)	(0.0006)
	[39.68]	[-320.83]	[42.38]	[-212.38]	[15.65]	[46.41]
Returns in Test Scores	0.0180***	0.0023**	0.0171***	0.0022*	0.0185***	-0.0001
	(0.0021)	(0.0009)	(0.0021)	(0.0009)	(0.0015)	(0.0005)
	[24.10]	[-20.79]	[26.05]	[-13.52]	[22.96]	[-0.55]
Levels in Controls	0.0190***	0.0044	0.0178***	0.0049	0.0207***	0.0032*
	(0.0025)	(0.0025)	(0.0025)	(0.0025)	(0.0018)	(0.0013)
	[25.42]	[-39.16]	[27.11]	[-29.62]	[25.67]	[13.37]
Residual	0.0080*	-0.0543***	0.0029	-0.0583***	0.0288***	0.0098***
	(0.0040)	(0.0050)	(0.0039)	(0.0049)	(0.0026)	(0.0027)
	[10.76]	[480.96]	[4.43]	[355.58]	[35.79]	[40.77]
Cohort FEs	✓	✓	✓	✓	✓	✓
School FEs	✓	✓	✓	✓	✓	✓
Census Tract FEs	✓	✓	✓	✓	✓	✓
Family Characteristics	✓	✓	✓	✓	✓	✓
White Offending Rate	.3332	.244	.3222	.2391	.0953	.0382
Black Offending Rate	.4079	.2327	.388	.2227	.1758	.0623
N	118898	120019	118898	120019	118898	120019

Notes: This table reports the estimates of each term of the Blinder-Oaxaca decomposition from Equation (2.3) for men separately by students' socio-economic status. Estimates were made from OLS estimates of the pooled model from Equation (2.2). Columns (1), (3), and (5) restrict our sample to economically disadvantaged students, and Columns (2), (4), and (6) only include economically non-disadvantaged students. Columns (1)-(2) decompose the gap in any offending by age 20, Columns (3)-(4) decompose the gap in any misdemeanor charges, and Columns (5)-(6) decompose the gap in felony charges. We do this decomposition using all math and reading scores from grades (3-8). Samples were restricted to individuals with observed test scores. Each row represents a different term in the breakdown. The first row represents the portion explained by average differences in test scores when using the white returns to test scores. The second row illustrates the difference in returns when using average test scores of black students test scores, the third row displays differences in S, and the fourth row displays the residual gap for the average student. The standard errors are in parentheses were calculated applying the delta method with robust standard error regression coefficients, assuming non-stochastic regressors (a la Oaxaca and Ransom (1998)). Each term's percent share of the raw black-white offending gap is reported below the standard errors in brackets.

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

Table 2.21: Blinder-Oaxaca Decomposition of Eighth Grade Math Test Scores Only and the Black-White Crime Gap at Age 20

	Any Charge	Misdemeanor Charge	Felony Charge
	(1)	(2)	(3)
Levels in Test Scores	0.0382***	0.0368***	0.0115***
	(0.0006)	(0.0006)	(0.0003)
	[47.79]	[50.14]	[20.88]
Returns in Test Scores	0.0104***	0.0101***	0.0089***
	(0.0006)	(0.0005)	(0.0003)
	[13.06]	[13.72]	[16.14]
Levels in Controls	0.0424***	0.0401***	0.0241***
	(0.0011)	(0.0011)	(0.0006)
	[53.02]	[54.64]	[43.58]
Residual	-0.0111***	-0.0136***	0.0107***
	(0.0016)	(0.0015)	(0.0009)
	[-13.85]	[-18.49]	[19.43]
Cohort FEs	✓	✓	✓
School FEs	✓	✓	✓
Census Tract FEs	✓	✓	✓
Family Characteristics	✓	✓	✓
White Offending Rate	.2106	.2055	.0374
Black Offending Rate	.2906	.2788	.0927
N	635218	635218	635218

Notes: This table reports the estimates of each term of the Blinder-Oaxaca (B-O) decomposition of the black-white offending gap for men, as depicted in Equation (2.3). These estimates differ from our other B-O decompositions, as we only use eighth grade math tests. Estimates were made from OLS estimates of the pooled model from Equation (2.2). Column (1) decomposes the gap in any offending by age 20, Column (2) decomposes the gap in any misdemeanor charges, and Column (3) analyzes the gap in felony charges. Samples were restricted to individuals with observed eighth grade math test scores. Each row represents a different term in the breakdown. The first row represents the portion explained by average differences in test scores when using the white returns to test scores. The second row illustrates the difference in returns when using average test scores of black students test scores, the third row displays differences in S, and the fourth row displays the residual gap for the average student. The standard errors are in parentheses were calculated applying the delta method with robust standard error regression coefficients, assuming non-stochastic regressors (a la Oaxaca and Ransom (1998)). Each term's percent share of the raw black-white offending gap is reported below the standard errors in brackets.

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

Chapter 3

Do Parents Value School Effectiveness?

3.1 Introduction

Recent education reforms in the United States, including charter schools, school vouchers, and district-wide open enrollment plans, increase parents' power to choose schools for their children. School choice allows households to avoid undesirable schools and forces schools to satisfy parents' preferences or risk losing enrollment. Proponents of choice argue that this competitive pressure is likely to generate system-wide increases in school productivity and boost educational outcomes for students (Friedman, 1962; Chubb and Moe, 1990; Hoxby, 2003). By decentralizing school quality assessment and allowing parents to act on local information, school choice may provide better incentives for educational effectiveness than could be achieved by a centralized accountability system. Choice may also improve outcomes by allowing students to sort into schools that suit their particular educational needs, resulting in improved match quality (Hoxby, 2000). These arguments have motivated recent policy efforts to expand school choice (e.g., DeVos, 2017).

If choice is to improve educational effectiveness, parents' choices must result in rewards for effective schools and sanctions for ineffective ones. Our use of the term "effective" follows Rothstein (2006): an effective school is one that generates causal improvements in student outcomes. Choice need not improve school effectiveness if it is not the basis for how parents choose between schools. For example, parents may value attributes such as facilities, convenience, student satisfication, or peer composition in a manner that does not align with educational impacts (Hanushek, 1981; Jacob and Lefgren, 2007). Moreover, while models in which parents value schools according to their effectiveness are an important benchmark in the academic literature (e.g., Epple, Figlio, and Romano, 2004), it may be difficult for parents to separate a school's effectiveness from the composition of its student body (Kane and Staiger, 2002). If parent choices reward schools that recruit higher-achieving students rather than schools that improve outcomes, school choice may increase resources devoted to screening and selection rather than better instruction (Ladd, 2002; MacLeod and Urquiola, 2015). Consistent with these possibilities, Rothstein (2006) shows that cross-district relationships among school choice, sorting patterns, and student outcomes fail to match the predictions of a model in which school effectiveness is the primary determinant of parent preferences.

This paper offers new evidence on the links between parent preferences, school effectiveness, and peer quality based on choice and outcome data for more than 250,000 applicants in New York City's centralized high school assignment mechanism. Each year, thousands of New York City high school applicants rank-order schools, and the mechanism assigns students to schools using the deferred acceptance (DA) algorithm (Gale and Shapley, 1962; Abdulkadiroğlu, Pathak, and Roth, 2005). The DA mechanism is strategy-proof: truthfully ranking schools is a weakly dominant strategy for students (Dubins and Freedman, 1981; Roth, 1982). This fact motivates our assumption that applicants' rankings measure their true preferences for schools.[1] We summarize these preferences by fitting discrete choice models to applicants' rank-ordered preference lists.

We then combine the preference estimates with estimates of school treatment effects on test scores, high school graduation, college attendance, and college choice. Treatment effect estimates come from "value-added" regression models of the sort commonly used to measure causal effects of teachers and schools (Todd and Wolpin, 2003; Koedel, Mihaly, and Rockoff, 2015). We generalize the conventional value-added approach to allow for match effects in academic outcomes and to relax the selection-on-observables assumption underlying standard models. Recent evidence suggests that value-added models controlling only for observables provide quantitatively useful but biased estimates of causal effects due to selection on unobservables (Rothstein, 2010, 2017; Chetty et al., 2014a; Angrist et al., 2017). We therefore use the rich information on preferences contained in students' rank-ordered choice lists to correct our estimates for selection on unobservables. This selection correction is implemented by extending the classic multinomial logit control function estimator of Dubin and McFadden (1984) to a setting where rankings of multiple alternatives are known.

The final step of our analysis relates the choice model and treatment effect estimates to measure preferences for school effectiveness. The choice and outcome models we estimate allow preferences and causal effects to vary flexibly with student characteristics. Our specifications accommodate the possibility that schools are more effective for specific types of students and that applicants choose schools that are a good match for their student type. We compare the degree to which parent preferences are explained by overall school effectiveness, match quality, and peer quality, defined as the component of a school's average outcome due to selection rather than effectiveness.

[1] As we discuss in Section 3.2, DA is strategy-proof when students are allowed to rank every school, but the New York City mechanism only allows applicants to rank 12 choices. Most students do not fill their preference lists, however, and truthful ranking is a dominant strategy in this situation (Haeringer and Klijn, 2009; Pathak and Sönmez, 2013). Fack, Grenet, and He (2015) propose empirical approaches to measuring student preferences without requiring that truth-telling is the unique equilibrium.

We find preferences are positively correlated with both peer quality and causal effects on student outcomes. More effective schools enroll higher-ability students, however, and preferences are unrelated to school effectiveness after controlling for peer quality. We also find little evidence of selection on match effects: on balance, parents do not prefer schools that are especially effective for their own children, and students do not enroll in schools that are a better-than-average match. These patterns are similar for short-run achievement test scores and longer-run postsecondary outcomes. Looking across demographic and baseline achievement groups, we find no evidence that any subgroup places positive weight on school effectiveness once we adjust for peer quality.

These findings do not indicate that parents choose schools irrationally; they may use peer characteristics to proxy for school effectiveness if the latter is difficult to observe, or value peer quality independently of impacts on academic outcomes. Regardless of the mechanism, however, our results imply that parents' choices penalize schools that enroll low achievers rather than schools that offer poor instruction. As a result, school choice programs may generate stronger incentives for screening and selection than for improved academic quality. We provide suggestive evidence that schools have responded to these incentives by increasing screening since the introduction of centralized assignment in New York City.

Our analysis complements Rothstein's (2006) indirect test with a direct assessment of the relationships among parent preferences, peer quality, and school effectiveness based on unusually rich choice and outcome data. The results also contribute to a large literature studying preferences for school quality (Black, 1999; Figlio and Lucas, 2004; Bayer, Ferreira, and McMillan, 2007; Hastings and Weinstein, 2008; Burgess, Greaves, Vignoles, and Wilson, 2014; Imberman and Lovenheim, 2016). These studies show that housing prices and household choices respond to school performance levels, but they do not typically separate responses to causal school effectiveness and peer quality. Our findings are also relevant to theoretical and empirical research on the implications of school choice for sorting and stratification (Epple and Romano, 1998; Epple et al., 2004; Hsieh and Urquiola, 2006; Barseghyan, Clark, and Coate, 2014; Altonji, Huang, and Taber, 2015; Avery and Pathak, 2015; MacLeod and Urquiola, 2015; MacLeod, Riehl, Saavedra, and Urquiola, 2017). In addition, our results help to reconcile some surprising findings from recent studies of school choice. Cullen, Jacob, and Levitt (2006) find limited achievement effects of admission to preferred schools in Chicago, while Walters (2018) documents that disadvantaged students in Boston are less likely to apply to charter schools than more advantaged students despite experiencing larger achievement benefits. Angrist, Pathak., and Walters (2013) and Abdulkadiroğlu, Pathak, and Walters (2018) report on two settings where parents opt for schools that reduce student achievement. These patterns are consistent with our finding that school choices are not driven by school effectiveness.

Finally, our analysis adds to a recent series of studies leveraging preference data from centralized school assignment mechanisms to investigate school demand (Hastings, Kane, and Staiger, 2009; Harris and Larsen, 2014; Fack et al., 2015; Abdulkadiroğlu, Agarwal, and Pathak, 2017a; Glazerman and Dotter, 2016; Kapor, Neilson, and Zimmerman, 2017; Agarwal and Somaini, 2018). Some of these studies analyze assignment mechanisms that provide incentives to strategically misreport preferences, while others measure academic quality using average test scores rather than distinguishing between peer quality and school effectiveness or looking at longer-run outcomes. We build on this previous work by using data from a strategy-proof mechanism to separately estimate preferences for peer quality and causal effects on multiple measures of academic success.

The rest of the paper is organized as follows. The next section describes school choice in New York City and the data used for our analysis. Section 3.3 develops a conceptual framework for analyzing school effectiveness and peer quality, and Section 3.4 details our empirical approach. Section 3.5 summarizes estimated distributions of student preferences and school treatment effects. Section 3.6 links preferences to peer quality and school effectiveness, and Section 3.7 discusses implications of these relationships. Section 3.8 concludes and offers some directions for future research.

3.2 Setting and Data

New York City High Schools

The New York City public school district annually enrolls roughly 90,000 ninth graders at more than 400 high schools. Rising ninth graders planning to attend New York City's public high schools submit applications to the centralized assignment system. Before 2003 the district used an uncoordinated school assignment process in which students could receive offers from more than one school. Motivated in part by insights derived from the theory of market design, in 2003 the city adopted a coordinated single-offer assignment mechanism based on the student-proposing deferred acceptance (DA) algorithm (Gale and Shapley, 1962; Abdulkadiroğlu et al., 2005; Abdulkadiroğlu, Pathak, and Roth, 2009). Abdulkadiroğlu et al. (2017a) show that introducing coordinated assignment reduced the share of administratively assigned students and likely improved average household welfare.

Applicants report their preferences for schooling options to the assignment mechanism by submitting rank-ordered lists of up to 12 academic programs. An individual school may operate more than one program. To aid families in their decisionmaking the New York City Department of Education (DOE) distributes a directory that provides an overview of the high school admission process, key dates, and an information page for each high school. A school's information page includes a brief statement of its mission, a list of offered programs, courses and extracurricular activities, pass rates on New York Regents standardized tests, and the school's graduation rate (New York City Department of Education, 2003). DOE also issues annual schools reports that list basic demographics, teacher characteristics, school expenditures, and Regents performance levels. During the time period of our study (2003-2007) these reports did not include measures of test score growth, though such measures have been added more recently (New York City Department of Education, 2004, 2017).

Academic programs prioritize applicants in the centralized admission system using a mix of factors. Priorities depend on whether a program is classified as unscreened, screened, or an educational option program. Unscreened programs give priority to students based on residential zones and (in some cases) to those who attend an information session. Screened programs use these factors and may also assign priorities based on prior grades, standardized test scores, and attendance. Educational option programs use screened criteria for some of their seats and unscreened criteria for the rest. Random numbers are used to order applicants with equal priority. A small group of selective high schools, including New York City's exam schools, admit students in a parallel system outside the main round of the assignment process (Abdulkadiroğlu, Angrist, and Pathak, 2014).

The DA algorithm combines student preferences with program priorities to generate a single program assignment for each student. In the initial step of the algorithm, each student proposes to her first-choice program. Programs provisionally accept students in order of priority up to capacity and reject the rest. In subsequent rounds, each student rejected in the previous step proposes to her most-preferred program among those that have not previously rejected her, and programs reject provisionally accepted applicants in favor of new applicants with higher priority. This process iterates until all students are assigned to a program or all unassigned students have been rejected by every program they have ranked. During our study time period, students left unassigned in the main round participate in a supplementary DA round in which they rank up to 12 additional programs with available seats. Any remaining students are administratively assigned by the district. About 82 percent, 8 percent, and 10 percent of applicants are assigned in the main, supplementary, and administrative rounds, respectively (Abdulkadiroğlu et al., 2017a).

An attractive theoretical property of the DA mechanism is that it is strategy-proof: since high-priority students can displace those with lower priority in later rounds of the process, listing schools in order of true preferences is a dominant strategy in the mechanism's canonical version. This property, however, requires students to have the option to rank all schools (Haeringer and Klijn, 2009; Pathak and Sönmez, 2013). As we show below, more than 70 percent of students rank fewer than 12 programs, meaning that truthful ranking of schools is a dominant strategy for the majority of applicants. The instructions provided with the New York City high school application also directly instruct students to rank schools in order of their true preferences (New York City Department of Education, 2003). In the analysis to follow, we therefore interpret students' rank-ordered lists as truthful reports of their preferences. We also probe the robustness of our findings to violations of this assumption by reporting results based on students that rank fewer than 12 choices.[2]

Data and Samples

The data used here are extracted from a DOE administrative information system covering all students enrolled in New York City public schools between the 2003-2004 and 2012-2013 school years. These data include school enrollment, student demographics, home addresses,

[2] Along similar lines, Abdulkadiroğlu et al. (2017a) show that preference estimates using only the top ranked school, the top three schools, and all but the last ranked school are similar.

scores on New York Regents standardized tests, Preliminary SAT (PSAT) scores, and high school graduation records, along with preferences submitted to the centralized high school assignment mechanism. A supplemental file from the National Student Clearinghouse (NSC) reports college enrollment for students graduating from New York City high schools between 2009 and 2012. A unique student identifier links records across these files.

We analyze high school applications and outcomes for four cohorts of students enrolled in New York City public schools in eighth grade between 2003-2004 and 2006-2007. This set of students is used to construct several samples for statistical analysis. The choice sample, used to investigate preferences for schools, consists of all high school applicants with baseline (eighth grade) demographic, test score, and address information. Our analysis of school effectiveness uses subsamples of the choice sample corresponding to each outcome of interest. These outcome samples include students with observed outcomes, baseline scores, demographics, and addresses, enrolled for ninth grade at one of 316 schools with at least 50 students for each outcome. The outcome samples also exclude students enrolled at the nine selective high schools that do not admit students via the main DA mechanism. Appendix 3.11 and Appendix Table 3.11 provide further details on data sources and sample construction.

Key outcomes in our analysis include Regents math standardized test scores, PSAT scores, high school graduation, college attendance, and college quality. The high school graduation outcome equals one if a student graduates within five years of her projected high school entry date given her eighth grade cohort. Likewise, college attendance equals one for students who enroll in any college (two or four year) within two years of projected on-time high school graduation. The college quality variable, derived from Internal Revenue Service tax record statistics reported by Chetty, Friedman, Saez, Turner, and Yagan (2017b), equals the mean 2014 income for children born between 1980 and 1982 who attended a student's college. The mean income for the non-college population is assigned to students who do not enroll in a college. While this metric does not distinguish between student quality and causal college effectiveness, it provides an accurate measure of the selectivity of a student's college. It has also been used elsewhere to assess effects of education programs on the intensive margin of college attendance (Chetty et al., 2011, 2014b). College attendance and quality are unavailable for the 2003-2004 cohort because the NSC data window does not allow us to determine whether students in this cohort were enrolled in college within two years of projected high school graduation.

Descriptive statistics for the choice and outcome samples appear in Table 3.1. These statistics show that New York City schools serve a disadvantaged urban population. Seventy-three percent of students are black or hispanic, and 65 percent are eligible for a subsidized lunch. Data from the 2011-2015 American Community Surveys shows that the average student in the choice sample lives in a census tract with a median household income of $50,136 in 2015 dollars. Observed characteristics are generally similar for students in the choice and outcome samples. The average PSAT score in New York City is 116, about one standard deviation below the US average (the PSAT is measured on a 240 point scale, normed to have a mean of 150 and a standard deviation of 30). The five-year high school graduation rate is 61 percent, and 48 percent of students attend some college within two years of graduation.

Choice Lists

New York City high school applicants tend to prefer schools near their homes, and most do not fill their choice lists. These facts are shown in Table 3.2, which summarizes rank-ordered preference lists in the choice sample. As shown in column (1), 93 percent of applicants submit a second choice, about half submit eight or more choices, and 28 percent submit the maximum 12 allowed choices. Column (2) shows that students prefer schools located in their home boroughs: 85 percent of first-choice schools are in the same borough as the student's home address, and the fraction of other choices in the home borough are also high. Abdulkadiroğlu et al. (2017a) report that for 2003-04, 193 programs restricted eligibility to applicants who reside in the same borough. The preference analysis to follow, therefore, treats schools in a student's home borough as her choice set and aggregates schools in other boroughs into a single outside option. Column (3), which reports average distances (measured as great-circle distance in miles) for each choice restricted to schools in the home borough, shows that students rank nearby schools higher within boroughs as well.

Applicants also prefer schools with strong academic performance. The last column of Table 3.2 reports the average Regents high school math score for schools at each position on the rank list. Regents scores are normalized to have mean zero and standard deviation one in the New York City population. To earn a high school diploma in New York state, students must pass a Regents math exam. These results reveal that higher-ranked schools enroll students with better math scores. The average score at a first-choice school is 0.2 standard deviations (σ) above the city average, and average scores monotonically decline with rank. PSAT, graduation, college enrollment, and college quality indicators also decline with rank. Students and parents clearly prefer schools with high achievement levels. Our objective in the remainder of this paper is to decompose this pattern into components due to preferences for school effectiveness and peer quality.

3.3 Conceptual Framework

Consider a population of students indexed by i, each of whom attends one of J schools. Let Y_{ij} denote the potential value of some outcome of interest for student i if she attends school j. The projection of Y_{ij} on a vector of observed characteristics, X_i, is written:

$$Y_{ij} = \alpha_j + X_i'\beta_j + \epsilon_{ij}, \tag{3.1}$$

where $E\left[\epsilon_{ij}\right] = E\left[X_i\epsilon_{ij}\right] = 0$ by definition of α_j and β_j. The coefficient vector β_j measures the returns to observed student characteristics at school j, while ϵ_{ij} reflects variation in potential outcomes unexplained by these characteristics. We further normalize $E\left[X_i\right] = 0$, so $\alpha_j = E\left[Y_{ij}\right]$ is the population mean potential outcome at school j. The realized outcome for student i is $Y_i = \sum_j 1\left\{S_i = j\right\} Y_{ij}$, where $S_i \in \{1...J\}$ denotes school attendance.

We decompose potential outcomes into components explained by student ability, school effectiveness, and idiosyncratic factors. Let $A_i \equiv (1/J)\sum_j Y_{ij}$ denote student i's general ability, defined as the average of her potential outcomes across all schools. This variable describes how the student would perform at the average school. Adding and subtracting A_i on the right-hand side of (3.1) yields:

$$Y_{ij} = \underbrace{\bar{\alpha} + X_i'\bar{\beta} + \bar{\epsilon}_i}_{A_i} + \underbrace{(\alpha_j - \bar{\alpha})}_{ATE_j} + \underbrace{X_i'(\beta_j - \bar{\beta}) + (\epsilon_{ij} - \bar{\epsilon}_i)}_{M_{ij}}, \tag{3.2}$$

where $\bar{\alpha} = (1/J)\sum_j \alpha_j$, $\bar{\beta} = (1/J)\sum_j \beta_j$, and $\bar{\epsilon}_i = (1/J)\sum_j \epsilon_{ij}$. Equation (3.2) shows that student i's potential outcome at school j is the sum of three terms: the student's general ability, A_i; the school's average treatment effect, ATE_j, defined as the causal effect of school j relative to an average school for an average student; and a match effect, M_{ij}, which reflects student i's idiosyncratic suitability for school j. Match effects may arise either because of an interaction between student i's observed characteristics and the extra returns to characteristics at school j (captured by $X_i'(\beta_j - \bar{\beta})$) or because of unobserved factors that make student i more or less suitable for school j (captured by $\epsilon_{ij} - \bar{\epsilon}_i$).

This decomposition allows us to interpret variation in observed outcomes across schools using three terms. The average outcome at school j is given by:

$$E\left[Y_i|S_i = j\right] = Q_j + ATE_j + E\left[M_{ij}|S_i = j\right]. \tag{3.3}$$

Here $Q_j \equiv E\left[A_i|S_i = j\right]$ is the average ability of students enrolled at school j, a variable we label "peer quality." The quantity $E\left[M_{ij}|S_i = j\right]$ is the average suitability of j's students for this particular school. In a Roy (1951)-style model in which students sort into schools on the basis of comparative advantage in the production of Y_i, we would expect this average match effect to be positive for all schools. Parents and students may also choose schools on the basis of peer quality Q_j, overall school effectiveness ATE_j, or the idiosyncratic match M_{ij} for various outcomes.

3.4 Empirical Methods

The goal of our empirical analysis is to assess the roles of peer quality, school effectiveness, and academic match quality in applicant preferences. Our analysis proceeds in three steps. We first use rank-ordered choice lists to estimate preferences, thereby generating measures of each school's popularity. Next, we estimate schools' causal effects on test scores, high school graduation, college attendance, and college choice. Finally, we combine these two sets of estimates to characterize the relationships among school popularity, peer quality, and causal effectiveness.

Estimating Preferences

Let U_{ij} denote student i's utility from enrolling in school j, and let $\mathcal{J} = \{1...J\}$ represent the set of available schools. We abstract from the fact that students rank programs rather than schools by ignoring repeat occurrences of any individual school on a student's choice list. U_{ij} may therefore be interpreted as the indirect utility associated with student i's favorite program at school j. The school ranked first on a student's choice list is

$$R_{i1} = \arg\max_{j \in \mathcal{J}} \ U_{ij},$$

while subsequent ranks satisfy

$$R_{ik} = \arg \max_{j \in \mathcal{J} \setminus \{R_{im}: m<k\}} U_{ij}, \ k > 1.$$

Student i's rank-order list is then $R_i = (R_{i1}...R_{i\ell(i)})'$, where $\ell(i)$ is the length of the list submitted by this student.

We summarize these preference lists by fitting random utility models with parameters that vary according to observed student characteristics. Student i's utility from enrolling in school j is modeled as:

$$U_{ij} = \delta_{c(X_i)j} - \tau_{c(X_i)} D_{ij} + \eta_{ij}, \tag{3.4}$$

where the function $c(X_i)$ assigns students to covariate cells based on the variables in the vector X_i, and D_{ij} records distance from student i's home address to school j. The parameter δ_{cj} is the mean utility of school j for students in covariate cell c, and τ_c is a cell-specific distance parameter or "cost." We include distance in the model because a large body of evidence suggests it plays a central role in school choices (e.g., Hastings et al., 2009 and Abdulkadiroğlu et al., 2017a). We model unobserved tastes η_{ij} as following independent extreme value type I distributions conditional on X_i and $D_i = (D_{i1}...D_{iJ})'$. Equation (3.4) is therefore a rank-ordered multinomial logit model (Hausman and Ruud, 1987).

The logit model implies the conditional likelihood of the rank list R_i is:

$$\mathcal{L}\left(R_i | X_i, D_i\right) = \prod_{k=1}^{\ell(i)} \frac{\exp\left(\delta_{c(X_i)R_{ik}} - \tau_{c(X_i)} D_{iR_{ik}}\right)}{\sum_{j \in \mathcal{J} \setminus \{R_{im}: m<k\}} \exp\left(\delta_{c(X_i)j} - \tau_{c(X_i)} D_{ij}\right)}.$$

We allow flexible heterogeneity in tastes by estimating preference models separately for 360 covariate cells defined by the intersection of borough, sex, race (black, hispanic, or other), subsidized lunch status, above-median census tract income, and terciles of the mean of eighth grade math and reading scores. This specification follows several recent studies that flexibly parametrize preference heterogeneity in terms of observable characteristics (e.g., Hastings, Hortačsu, and Syverson, 2017 and Langer, 2016). Students rarely rank schools outside their home boroughs, so covariate cells often include zero students ranking any given out-of-borough school. We therefore restrict the choice set $\mathcal{J}$ to schools located in the home borough and aggregate all other schools into an outside option with utility normalized to zero. Maximum likelihood estimation of the preference parameters produces a list of school mean utilities along with a distance coefficient for each covariate cell.

Estimating School Effectiveness

Our analysis of school effectiveness aims to recover the parameters of the potential outcome equations defined in Section 3.3. We take two approaches to estimating these parameters.

Approach 1: Selection on observables

The first set of estimates is based on the assumption:

$$E\left[Y_{ij}|X_i, S_i\right] = \alpha_j + X_i'\beta_j,\ j = 1...J. \tag{3.5}$$

This restriction, often labeled "selection on observables," requires school enrollment to be as good as random conditional on the covariate vector X_i, which includes sex, race, subsidized lunch status, the log of median census tract income, and eighth grade math and reading scores. Assumption (3.5) implies that an ordinary least squares (OLS) regression of Y_i on school indicators interacted with X_i recovers unbiased estimates of α_j and β_j for each school. This fully interacted specification is a multiple-treatment extension of the Oaxaca-Blinder (1973) treatment effects estimator (Kline, 2011).[3] By allowing school effectiveness to vary with student characteristics, we generalize the constant effects "value-added" approach commonly used to estimate the contributions of teachers and schools to student achievement (Koedel et al., 2015).

The credibility of the selection on observables assumption underlying value-added estimators is a matter of continuing debate (Rothstein, 2010, 2017; Kane, McCaffrey, and Staiger, 2013; Bacher-Hicks et al., 2014; Chetty et al., 2014a, 2016, 2017a; Guarino, Reckase, and Wooldridge, 2015). Comparisons to results from admission lotteries indicate that school value-added models accurately predict the impacts of random assignment but are not perfectly unbiased (Deming, 2014; Angrist, Hull, Pathak, and Walters, 2016b; Angrist et al., 2017). Selection on observables may also be more plausible for test scores than for longer-run outcomes, for which lagged measures of the dependent variable are not available (Chetty et al., 2014a). We therefore report OLS estimates as a benchmark and compare these to estimates from a more general strategy that relaxes assumption (3.5).

Approach 2: Rank-ordered control functions

Our second approach is motivated by the restriction:

$$E\left[Y_{ij}|X_i, D_i, \eta_{i1}...\eta_{iJ}, S_i\right] = \alpha_j + X_i'\beta_j + g_j(D_i, \eta_{i1}, .., \eta_{iJ}),\ j = 1...J. \tag{3.6}$$

[3] We also include main effects of borough so that the model includes the same variables used to define covariate cells in the preference estimates.

This restriction implies that any omitted variable bias afflicting OLS value-added estimates is due either to spatial heterogeneity captured by distances to each school (D_i) or to the preferences underlying the rank-ordered lists submitted to the assignment mechanism (η_{ij}). The function $g_j(\cdot)$ allows potential outcomes to vary arbitrarily across students with different preferences over schools. Factors that lead students with the same observed characteristics, spatial locations, and preferences to ultimately enroll in different schools, such as school priorities, random rationing due to oversubscription, or noncompliance with the assignment mechanism, are presumed to be unrelated to potential outcomes.

Under assumption (3.6), comparisons of matched sets of students with the same covariates, values of distance, and rank-ordered choice lists recover causal effects of school attendance. This model is therefore similar to the "self-revelation" model proposed by Dale and Krueger (2002; 2014) in the context of postsecondary enrollment. Dale and Krueger assume that students reveal their unobserved "types" via the selectivity of their college application portfolios, so college enrollment is as good as random among students that apply to the same schools. Similarly, (3.6) implies that high school applicants reveal their types through the content of their rank-ordered preference lists.

Though intuitively appealing, full nonparametric matching on rank-ordered lists is not feasible in practice because few students share the exact same rankings. We therefore use the structure of the logit choice model in equation (3.4) to derive a parametric approximation to this matching procedure. Specifically, we replace equation (3.6) with the assumption:

$$E\left[Y_{ij}|X_i, D_i, \eta_{i1}...\eta_{iJ}, S_i\right] = \alpha_j + X_i'\beta_j + D_i'\gamma + \sum_{k=1}^{J}\psi_k \times (\eta_{ik} - \mu_\eta) + \varphi \times (\eta_{ij} - \mu_\eta),\ j = 1...J, \tag{3.7}$$

where $\mu_\eta \equiv E\left[\eta_{ij}\right]$ is Euler's constant.[4] As in the multinomial logit selection model of Dubin and McFadden (1984), equation (3.7) imposes a linear relationship between potential outcomes and the unobserved logit errors. Functional form assumptions of this sort are common in multinomial selection models with many alternatives, where requirements for nonparametric identification are very stringent (Lee, 1983; Dahl, 2002; Heckman, Urzua, and Vytlacil, 2008).[5]

Equation (3.7) accommodates a variety of forms of selection on unobservables. The coefficient ψ_k represents an effect of the preference for school k common to all potential outcomes. This permits students with strong preferences for particular schools to have higher or lower general ability A_i. The parameter φ captures an additional match effect of the preference for school j on the potential outcome at this specific school. The model therefore allows for "essential" heterogeneity linking preferences to unobserved match effects in student outcomes (Heckman, Urzua, and Vytlacil, 2006b). A Roy (1951)-style model of selection on gains would imply $\varphi > 0$, but we do not impose this restriction.

[4] The means of both X_i and D_i are normalized to zero to maintain the interpretation that $\alpha_j = E[Y_{ij}]$.

[5] As discussed in Section 3.6, we also estimate an alternative model that includes fixed effects for first choice schools.

By iterated expectations, equation (3.7) implies that mean observed outcomes at school j are:

$$E\left[Y_i|X_i, D_i, R_i, S_i = j\right] = \alpha_j + X_i'\beta_j + D_i'\gamma + \sum_{k=1}^{J} \psi_k\lambda_k\left(X_i, D_i, R_i\right) + \varphi\lambda_j(X_i, D_i, R_i), \quad (3.8)$$

where $\lambda_k\left(X_i, D_i, R_i\right) \equiv E\left[\eta_{ik} - \mu_\eta|X_i, D_i, R_i\right]$ gives the mean preference for school k conditional on a student's characteristics, spatial location, and preference list. The $\lambda_k(\cdot)$'s serve as "control functions" correcting for selection on unobservables (Heckman and Robb, 1985; Blundell and Matzkin, 2014; Wooldridge, 2015). As shown in Appendix 3.12, these functions are generalizations of the formulas derived by Dubin and McFadden (1984), extended to account for the fact that we observe a list of several ranked alternatives rather than just the most preferred choice.

Note that equation (3.8) includes main effects of distance to each school; we do not impose an exclusion restriction for distance. Identification of the selection parameters ψ_k and φ comes from variation in preference rankings for students who enroll at the same school conditional on covariates and distance. Intuitively, if students who rank school j highly do better than expected given their observed characteristics at all schools, we will infer that $\psi_j > 0$. If these students do better than expected at school j but not elsewhere, we will infer that $\varphi > 0$.

We use the choice model parameters to build first-step estimates of the control functions, then estimate equation (3.8) in a second-step OLS regression of Y_i on school indicators and their interactions with X_i, controlling for D_i and the estimated $\lambda_k(\cdot)$ functions.[6] We adjust inference for estimation error in the control functions via a two-step extension of the score bootstrap procedure of Kline and Santos (2012). As detailed in Appendix 3.12, the score bootstrap avoids the need to recalculate the first-step logit estimates or the inverse variance matrix of the second-step regressors in the bootstrap iterations.

The joint distribution of peer quality and school effectiveness

Estimates of equations (3.5) and (3.7) may be used to calculate each school's peer quality. A student's predicted ability in the value-added model is

$$\hat{A}_i = \frac{1}{J}\sum_{j=1}^{J}\left[\hat{\alpha}_j + X_i'\hat{\beta}_j\right], \quad (3.9)$$

[6] The choice model uses only preferences over schools in students' home boroughs, so $\lambda_k(\cdot)$ is undefined for students outside school k's borough. We therefore include dummies for missing values and code the control functions to zero for these students. We similarly code D_{ik} to zero for students outside of school k's borough and include borough indicators so that the distance coefficients are estimated using only within-borough variation. Our key results are not sensitive to dropping students attending out-of-borough schools from the sample.

where $\hat{\alpha}_j$ and $\hat{\beta}_j$ are OLS value-added coefficients. Predicted ability in the control function model adds estimates of the distance and control function terms in equation (3.8). Estimated peer quality at school j is then $\hat{Q}_j = \sum_i 1\{S_i = j\}\hat{A}_i / \sum_i 1\{S_i = j\}$, the average predicted ability of enrolled students.

The end result of our school quality estimation procedure is a vector of estimates for each school, $\hat{\theta}_j = (\hat{\alpha}_j, \hat{\beta}_j', \hat{Q}_j)'$. The vector of parameters for the control function model also includes an estimate of the selection coefficient for school j, $\hat{\psi}_j$. These estimates are unbiased but noisy measures of the underlying school-specific parameters θ_j. We investigate the distribution of θ_j using the following hierarchical model:

$$\begin{aligned} \hat{\theta}_j | \theta_j &\sim N(\theta_j, \Omega_j), \\ \theta_j &\sim N(\mu_\theta, \Sigma_\theta). \end{aligned} \tag{3.10}$$

Here Ω_j is the sampling variance of the estimator $\hat{\theta}_j$, while μ_θ and Σ_θ govern the distribution of latent parameters across schools. In a hierarchical Bayesian framework μ_θ and Σ_θ are hyperparameters describing a prior distribution for θ_j. We estimate these hyperparameters by maximum likelihood applied to model (3.10), approximating Ω_j with an estimate of the asymptotic variance of $\hat{\theta}_j$.[7] The resulting estimates of μ_θ and Σ_θ characterize the joint distribution of peer quality and school treatment effect parameters, purged of the estimation error in $\hat{\theta}_j$.

This hierarchical model can also be used to improve estimates of parameters for individual schools. An empirical Bayes (EB) posterior mean for θ_j is given by

$$\theta_j^* = \left(\hat{\Omega}_j^{-1} + \hat{\Sigma}_\theta^{-1}\right)^{-1} \left(\hat{\Omega}_j^{-1}\hat{\theta}_j + \hat{\Sigma}_\theta^{-1}\hat{\mu}_\theta\right),$$

where $\hat{\Omega}_j$, $\hat{\mu}_\theta$ and $\hat{\Sigma}_\theta$ are estimates of Ω_j, μ_θ and Σ_θ. Relative to the unbiased but noisy estimate $\hat{\theta}_j$, this EB shrinkage estimator uses the prior distribution to reduce sampling variance at the cost of increased bias, yielding a minimum mean squared error (MSE) prediction of θ_j (Robbins, 1956; Morris, 1983). This approach parallels recent work applying shrinkage methods to estimate causal effects of teachers, schools, neighborhoods, and hospitals (Chetty et al., 2014a; Hull, 2016; Angrist et al., 2017; Chetty and Hendren, 2017; Finkelstein, Gentzkow, Hull, and Williams, 2017). Appendix 3.12 further describes our EB estimation strategy. In addition to reducing MSE, empirical Bayes shrinkage eliminates attenuation bias that would arise in models using elements of $\hat{\theta}_j$ as regressors (Jacob and Lefgren, 2008). We exploit this property by regressing estimates of school popularity on EB posterior means in the final step of our empirical analysis.

[7] The peer quality estimates $\hat{Q}_j$ are typically very precise, so we treat peer quality as known rather than estimated when fitting the hierarchical model.

Linking Preferences to School Effectiveness

We relate preferences to peer quality and causal effects with regressions of the form:

$$\hat{\delta}_{cj} = \kappa_c + \rho_1 Q_j^* + \rho_2 ATE_j^* + \rho_3 M_{cj}^* + \xi_{cj}, \tag{3.11}$$

where $\hat{\delta}_{cj}$ is an estimate of the mean utility of school j for students in covariate cell c, κ_c is a cell fixed effect, and Q_j^* and ATE_j^* are EB posterior mean predictions of peer quality and average treatment effects. The variable M_{cj}^* is an EB prediction of the mean match effect of school j for students in cell c. Observations in equation (3.11) are weighted by the inverse sampling variance of $\hat{\delta}_{cj}$. We use the variance estimator proposed by Cameron, Gelbach, and Miller (2011) to double-cluster inference by cell and school. Two-way clustering accounts for correlated estimation errors in $\hat{\delta}_{cj}$ across schools within a cell as well as unobserved determinants of popularity common to a given school across cells.

We estimate equation (3.11) separately for Regents test scores, PSAT scores, high school graduation, college attendance, and college quality. The parameters ρ_1, ρ_2, and ρ_3 measure how preferences relate to peer quality, overall school effectiveness, and match quality.

3.5 Parameter Estimates

Preference Parameters

Table 3.3 summarizes the distribution of household preference parameters across the 316 high schools and 360 covariate cells in the choice sample. The first row reports estimated standard deviations of the mean utility δ_{cj} across schools and cells, while the second row displays the mean and standard deviation of the cell-specific distance cost τ_c. School mean utilities are deviated from cell averages to account for differences in the reference category across boroughs, and calculations are weighted by cell size. We adjust these standard deviations for sampling error in the estimated preference parameters by subtracting the average squared standard error from the sample variance of mean utilities.

Consistent with the descriptive statistics in Table 3.1, the preference estimates indicate that households dislike more distant schools. The mean distance cost is 0.33. This implies that increasing the distance to a particular school by one mile reduces the odds that a household prefers this school to another in the same borough by 33 percent. The standard deviation of the distance cost across covariate cells is 0.12. While there is significant heterogeneity in distastes for distance, all of the estimated distance costs are positive, suggesting that all subgroups prefer schools closer to home.

The estimates in Table 3.3 reveal significant heterogeneity in tastes for schools both within and between subgroups. The within-cell standard deviation of school mean utilities, which measures the variation in δ_{cj} across schools j for a fixed cell c, equals 1.12. This is equivalent to roughly 3.4 (1.12/0.33) miles of distance, implying that households are willing to travel substantial distances to attend more popular schools. The between-cell standard deviation, which measures variation in δ_{cj} across c for a fixed j, is 0.50, equivalent to about 1.5 (0.50/0.33) miles of distance. The larger within-cell standard deviation indicates that students in different subgroups tend to prefer the same schools.

School Effectiveness and Peer Quality

Our estimates of school treatment effects imply substantial variation in both causal effects and sorting across schools. Table 3.4 reports estimated means and standard deviations of peer quality Q_j, average treatment effects ATE_j, and slope coefficients β_j. We normalize the means of Q_j and ATE_j to zero and quantify the variation in these parameters relative to the average school. As shown in column (2), the value-added model produces standard deviations of Q_j and ATE_j for Regents math scores equal to 0.29σ. This is somewhat larger than corresponding estimates of variation in school value-added from previous studies (usually around $0.15 - 0.2\sigma$; see, e.g., Angrist et al., 2017). One possible reason for this difference is that most students in our sample attend high school for two years before taking Regents math exams, while previous studies look at impacts after one year.

As shown in columns (3) and (4) of Table 3.4, the control function model attributes some of the variation in Regents math value-added parameters to selection bias. Adding controls for unobserved preferences and distance increases the estimated standard deviation of Q_j to 0.31σ and reduces the estimated standard deviation of ATE_j to 0.23σ. Figure 3.1, which compares value-added and control function estimates for all five outcomes, demonstrates that this pattern holds for other outcomes as well: adjusting for selection on unobservables compresses the estimated distributions of treatment effects. This compression is more severe for high school graduation, college attendance, and college quality than for Regents math and PSAT scores. Our findings are therefore consistent with previous evidence that bias in OLS value-added models is more important for longer-run and non-test score outcomes (see, e.g., Chetty et al., 2014b).

The bottom rows of Table 3.4 show evidence of substantial treatment effect heterogeneity across students. For example, the standard deviation of the slope coefficient on a black indicator equals 0.12σ in the control function model. This implies that holding the average treatment effect ATE_j fixed, a one standard deviation improvement in a school's match quality for black students boosts scores for these students by about a tenth of a standard deviation relative to whites. We also find significant variation in slope coefficients for gender (0.06σ), hispanic (0.11σ), subsidized lunch status (0.05σ), the log of median census tract income (0.05σ), and eighth grade math and reading scores (0.11σ and 0.05σ). The final row of column (3) reports a control function estimate of φ, the parameter capturing matching between unobserved preferences and Regents scores. This estimate indicates a positive relationship between preferences and the unobserved component of student-specific test score gains, but the magnitude of the coefficient is very small.[8]

[8] The average predicted value of $(\eta_{ij} - \mu_\eta)$ for a student's enrolled school in our sample is 2.0. Our estimate of φ therefore implies that unobserved match effects increase average test scores by about one percent of a

Our estimates imply that high-ability students tend to enroll in more effective schools. Table 3.5 reports correlations between Q_j and school treatment effect parameters based on control function estimates for Regents math scores. Corresponding value-added estimates appear in Appendix Table 3.12. The estimated correlation between peer quality and average treatment effects is 0.59. This may reflect either positive peer effects or higher-achieving students' tendency to enroll in schools with better inputs. Our finding that schools with high-ability peers are more effective contrasts with recent studies of exam schools in New York City and Boston, which show limited treatment effects for highly selective public schools (Abdulkadiroğlu et al., 2014; Dobbie and Fryer, 2014). Within the broader New York public high school system, we find a strong positive association between school effectiveness and average student ability.

Table 3.4 also reports estimated correlations of Q_j and ATE_j with the elements of the slope coefficient vector β_j. Schools with larger average treatment effects tend to be especially good for girls: the correlation between ATE_j and the female slope coefficient is positive and statistically significant. This is consistent with evidence from Deming, Hastings, Kane, and Staiger (2014) showing that girls' outcomes are more responsive to school value-added. We estimate a very high positive correlation between black and hispanic coefficients, suggesting that match effects tend to be similar for these two groups.

The slope coefficient on eighth grade reading scores is negatively correlated with peer quality and the average treatment effect. Both of these estimated correlations are below -0.4 and statistically significant. In other words, schools that enroll higher-ability students and produce larger achievement gains are especially effective at teaching low-achievers. In contrast to our estimate of the parameter φ, this suggests negative selection on the observed component of match effects in student achievement. Section 3.6 presents a more systematic investigation of this pattern by documenting the net relationship between preferences and treatment effects combining all student characteristics.

Patterns of estimates for PSAT scores, high school graduation, college attendance, and college quality are generally similar to results for Regents math scores. Appendix Tables 3.13-3.16 present estimated distributions of peer quality and school effectiveness for these longer-run outcomes. For all five outcomes, we find substantial variation in peer quality and average treatment effects, a strong positive correlation between these variables, and significant effect heterogeneity with respect to student characteristics. Overall, causal effects for the longer-run outcomes are highly correlated with effects on Regents math scores. This is evident in Figure 3.2, which plots EB posterior mean predictions of average treatment effects on Regents scores against corresponding predictions for the other four outcomes. These results are consistent with recent evidence that short-run test score impacts reliably predict effects on longer-run outcomes (Chetty et al., 2011; Dynarski, Hyman, and Schanzenbach, 2013; Angrist, Cohodes, Dynarski, Pathak, and Walters, 2016a).

standard deviation ($0.006\sigma \times 2.0 = 0.012\sigma$).

Decomposition of School Average Outcomes

We summarize the joint distribution of peer quality and school effectiveness by implementing the decomposition introduced in Section 3.3. Table 3.6 uses the control function estimates to decompose variation in school averages for each outcome into components explained by peer quality, school effectiveness, average match effects, and covariances of these components.

Consistent with the estimates in Table 3.4, both peer quality and school effectiveness play roles in generating variation in school average outcomes, but peer quality is generally more important. Peer quality explains 47 percent of the variance in average Regents scores (0.093/0.191), while average treatment effects explain 28 percent (0.054/0.191). The explanatory power of peer quality for other outcomes ranges from 49 percent (PSAT scores) to 83 percent (high school graduation), while the importance of average treatment effects ranges from 10 percent (PSAT scores) to 19 percent (log college quality).

Despite the significant variation in slope coefficients documented in Table 3.4, match effects are unimportant in explaining dispersion in school average outcomes. The variance of match effects accounts for only five percent of the variation in average Regents scores, and corresponding estimates for the other outcomes are also small. Although school treatment effects vary substantially across subgroups, there is not much sorting of students to schools on this basis, so the existence of potential match effects is of little consequence for realized variation in outcomes across schools.

The final three rows of Table 3.6 quantify the contributions of covariances among peer quality, treatment effects, and match effects. As a result of the positive relationship between peer quality and school effectiveness, the covariance between Q_j and ATE_j substantially increases cross-school dispersion in mean outcomes. The covariances between match effects and the other variance components are negative. This indicates that students at highly effective schools and schools with higher-ability students are less appropriately matched on the heterogeneous component of treatment effects, slightly reducing variation in school average outcomes.

3.6 Preferences, Peer Quality, and School Effectiveness

Productivity vs. Peers

The last step of our analysis compares the relative strength of peer quality and school effectiveness as predictors of parent preferences. Table 3.7 reports estimates of equation (3.11) for Regents math scores, first including Q_j^* and ATE_j^* one at a time and then including both variables simultaneously. Mean utilities, peer quality, and treatment effects are scaled in standard deviations of their respective school-level distributions, so the estimates can be interpreted as the standard deviation change in mean utility associated with a one standard deviation increase in Q_j or ATE_j.

Bivariate regressions show that school popularity is positively correlated with both peer quality and school effectiveness. Results based on the OLS value-added model, reported in columns (1) and (2), imply that a one standard deviation increase in Q_j is associated with a 0.42 standard deviation increase in mean utility, while a one standard deviation increase in ATE_j is associated with a 0.24 standard deviation increase in mean utility. The latter result contrasts with studies reporting no average test score impact of attending preferred schools (Cullen et al., 2006; Hastings et al., 2009). These studies rely on admission lotteries that shift relatively small numbers of students across a limited range of schools. Our results show that looking across all high schools in New York City, more popular schools tend to be more effective on average.

While preferences are positively correlated with school effectiveness, however, this relationship is entirely explained by peer quality. Column (3) shows that when both variables are included together, the coefficient on peer quality is essentially unchanged, while the coefficient on the average treatment effect is rendered small and statistically insignificant. The ATE_j coefficient also remains precise: we can rule out increases in mean utility on the order of 0.06 standard deviations associated with a one standard deviation change in school value-added at conventional significance levels. The control function estimates in columns (5)-(7) are similar to the value-added estimates; in fact, the control function results show a small, marginally statistically significant negative association between school effectiveness and popularity after controlling for peer quality.

Columns (4) and (8) of Table 3.7 explore the role of treatment effect heterogeneity by adding posterior mean predictions of match quality to equation (3.11), also scaled in standard deviation units of the distribution of match effects across schools and cells. The match coefficient is negative for both the value-added and control function models, and the control function estimate is statistically significant. This reflects the negative correlation between baseline test score slope coefficients and peer quality reported in Table 3.5: schools that are especially effective for low-achieving students tend to be more popular among high-achievers and therefore enroll more of these students despite their lower match quality. This is consistent with recent studies of selection into early-childhood programs and charter schools, which also find negative selection on test score match effects (Cornelissen, Dustmann, Raute, and Schönberg, 2016; Kline and Walters, 2016; Walters, 2018).

Figure 3.3 presents a graphical summary of the links among preferences, peer quality, and treatment effects by plotting bivariate and multivariate relationships between mean utility (averaged across covariate cells) and posterior predictions of Q_j and ATE_j from the control function model. Panel A shows strong positive bivariate correlations for both variables. Panel B plots mean utilities against residuals from a regression of Q_j^* on ATE_j^* (left-hand panel) and residuals from a regression of ATE_j^* on Q_j^* (right-hand panel). Adjusting for school effectiveness has little effect on the relationship between preferences and peer quality. In contrast, partialing out peer quality eliminates the positive association between popularity and effectiveness.

Preferences and Effects on Longer-run Outcomes

Parents may care about treatment effects on outcomes other than short-run standardized test scores. We explore this by estimating equation (3.11) for PSAT scores, high school graduation, college attendance, and log college quality.

Results for these outcomes are similar to the findings for Regents math scores: preferences are positively correlated with average treatment effects in a bivariate sense but are uncorrelated with treatment effects conditional on peer quality. Table 3.8 reports results based on control function estimates of treatment effects. The magnitudes of all treatment effect coefficients are small, and the overall pattern of results suggests no systematic relationship between preferences and school effectiveness conditional on peer composition. We find a modest positive relationship between preferences and match effects for log college quality, but corresponding estimates for PSAT scores, high school graduation, and college attendance are small and statistically insignificant. This pattern contrasts with results for the Norwegian higher education system, reported by Kirkebøen, Leuven, and Mogstad (2016), which show sorting into fields of study based on heterogeneous earnings gains. Unlike Norwegian college students, New York City's high school students do not prefer schools with higher academic match quality.

Heterogeneity in Preferences for Peer and School Quality

Previous evidence suggests that parents of higher-income, higher-achieving students place more weight on academic performance levels when choosing schools (Hastings et al., 2009). This pattern may reflect either greater responsiveness to peer quality or more sensitivity to causal school effectiveness. If parents of high-achievers value school effectiveness, choice may indirectly create incentives for schools to improve because better instruction will attract high-ability students, raising peer quality and therefore demand from other households. In Table 3.9 we investigate this issue by estimating equation (3.11) separately by sex, race, subsidized lunch status, and baseline test score category.

We find that no subgroup of households responds to causal school effectiveness. Consistent with previous work, we find larger coefficients on peer quality among non-minority students, richer students (those ineligible for subsidized lunches), and students with high baseline achievement. We do not interpret this as direct evidence of stronger preferences for peer ability among higher-ability students; since students are more likely to enroll at schools they rank highly, any group component to preferences will lead to a positive association between students' rankings and the enrollment share of others in the same group.[9] The key pattern in Table 3.9 is that, among schools with similar peer quality, no group prefers schools with greater causal impacts on academic achievement.

[9] This is a version of the "reflection problem" that plagues econometric investigations of peer effects (Manski, 1993).

Alternative Specifications

We investigate the robustness of our key results by estimating a variety of alternative specifications, reported in Appendix Tables 3.17-3.19. To assess the sensitivity of our estimates to reasonable changes in our measure of school popularity, Appendix Table 3.17 displays results from models replacing $\hat{\delta}_{cj}$ in equation (3.11) with the log share of students in a cell ranking a school first or minus the log sum of ranks in the cell (treating unranked schools as tied). These alternative measures of demand produce very similar results to the rank-ordered logit results in Table 3.7.

Estimates based on students' submitted rankings may not accurately describe demand if students strategically misreport their preferences in response to the 12-choice constraint on list length. As noted in Section 3.2, truthful reporting is a dominant strategy for the 72 percent of students that list fewer than 12 choices. Appendix Table 3.18 reports results based on rank-ordered logit models estimated in the subsample of unconstrained students. Results here are again similar to the full sample estimates, suggesting that strategic misreporting is not an important concern in our setting.

Equation (3.8) parameterizes the relationship between potential outcomes and preference rankings through the control functions $\lambda_k(\cdot)$. Columns (1)-(4) of Appendix Table 3.19 present an alternative parameterization that replaces the control functions with fixed effects for first choice schools. This approach ignores information on lower-ranked schools but more closely parallels the application portfolio matching approach in Dale and Krueger (2002; 2014). As a second alternative specification, columns (5)-(8) report estimates from a control function model that drops the distance control variables from equation (3.8). This model relies on an exclusion restriction for distance, a common identification strategy in the literature on educational choice (Card, 1995; Neal, 1997; Booker, Sass, Gill, and Zimmer, 2011; Walters, 2018; Mountjoy, 2017). These alternative approaches to estimating school effectiveness produce no meaningful changes in the results.

3.7 Discussion

The findings reported here inform models of school choice commonly considered in the literature. Theoretical analyses often assume parents know students' potential achievement outcomes and choose between schools on this basis. For example, Epple et al. (2004) and Epple and Romano (2008) study models in which parents value academic achievement and consumption of other goods, and care about peer quality only insofar as it produces higher achievement through peer effects. Hoxby (2000) argues that school choice may increase achievement by allowing students to sort on match quality. Such models imply that demand should be positively correlated with both average treatment effects and match effects conditional on peer quality, a prediction that is inconsistent with the pattern in Table 3.7.

Parents may choose between schools based on test score levels rather than treatment effects. Cullen et al. (2006) suggest confusion between levels and gains may explain limited effects of admission to preferred schools in Chicago. Since our setting has substantial variation in both levels and value-added, we can more thoroughly investigate this model of parent decision-making. If parents choose between schools based on average outcomes, increases in these outcomes due to selection and causal effectiveness should produce equal effects on popularity. In contrast, we find that demand only responds to the component of average outcomes that is due to enrollment of higher-ability students. That is, we can reject the view that parental demand is driven by performance levels: demand places no weight on the part of performance levels explained by value-added but significant weight on the part explained by peer quality.

It is important to note that our findings do not imply parents are uninterested in school effectiveness. Without direct information about treatment effects, for example, parents may use peer characteristics as a proxy for school quality, as in MacLeod and Urquiola (2015). In view of the positive correlation between peer quality and school effectiveness, this is a reasonable strategy for parents that cannot observe treatment effects and wish to choose effective schools. Effectiveness varies widely conditional on peer quality, however, so parents make substantial sacrifices in academic quality by not ranking schools based on effectiveness. Table 3.10 compares Regents math effects for observed preference rankings vs. hypothetical rankings in which parents order schools according to their effectiveness. The average treatment effect of first-choice schools would improve from 0.07σ to 0.43σ if parents ranked schools based on effectiveness, and the average match effect would increase from -0.04σ to 0.16σ. This implies that the average student loses more than half a standard deviation in math achievement by enrolling in her first-choice school rather than the most effective option.

The statistics in Table 3.10 suggest that if information frictions prevent parents from ranking schools based on effectiveness, providing information about school effectiveness could alter school choices considerably. These changes may be particularly valuable for disadvantaged students. As shown in Appendix Table 3.20, gaps in effectiveness between observed first-choice schools and achievement-maximizing choices are larger for students with lower baseline achievement. This is driven by the stronger relationship between peer quality and preferences for more-advantaged parents documented in Table 3.9. These results suggest reducing information barriers could lead to differential increases in school quality for disadvantaged students and reduce inequality in student achievement. On the other hand, the patterns documented here may also reflect parents' valuation of school amenities other than academic effectiveness rather than a lack of information about treatment effects.

Regardless of why parents respond to peer quality rather than school effectiveness, our results have important implications for the incentive effects of school choice programs. Since parents only respond to the component of school average outcomes that can be predicted by the ability of enrolled students, our estimates imply a school wishing to boost its popularity must recruit better students; improving outcomes by increasing causal effectiveness for a fixed set of students will have no impact on parent demand. Our results therefore suggest that choice may create incentives for schools to invest in screening and selection.

The evolution of admissions criteria used at New York City's high schools is consistent with the implication that schools have an increased incentive to screen applicants due to parents' demand for high-ability peers. After the first year of the new assignment mechanism, several school programs eliminated all lottery-based admissions procedures and became entirely screened. In the 2003-04 high school brochure, 36.8 percent of programs are screened, and this fraction jumps to 40.3 percent two years later. The Beacon High School in Manhattan, for example, switched from a school where half of the seats were assigned via random lottery in 2003-04 to a screened school the following year, where admissions is based on test performance, an interview and a portfolio of essays. Leo Goldstein High School for Sciences in Brooklyn underwent a similar transition. Both high schools frequent lists of New York City;s best public high schools (Linge and Tanzer, 2016). Compared to the first years of the new system, there has also been growth in the number of limited unscreened programs, which use a lottery but also give priority to students who attend an open house or high school fair. Compared to unscreened programs, prioritizing applicants who attend an information session provides an ordeal that favors applicants with time and resources thus resulting in positive selection (Disare, 2017). The number of limited unscreened programs nearly doubled from 106 to 210 from 2005 to 2012 (Nathanson, Corcoran, and Baker-Smith, 2013).

3.8 Conclusion

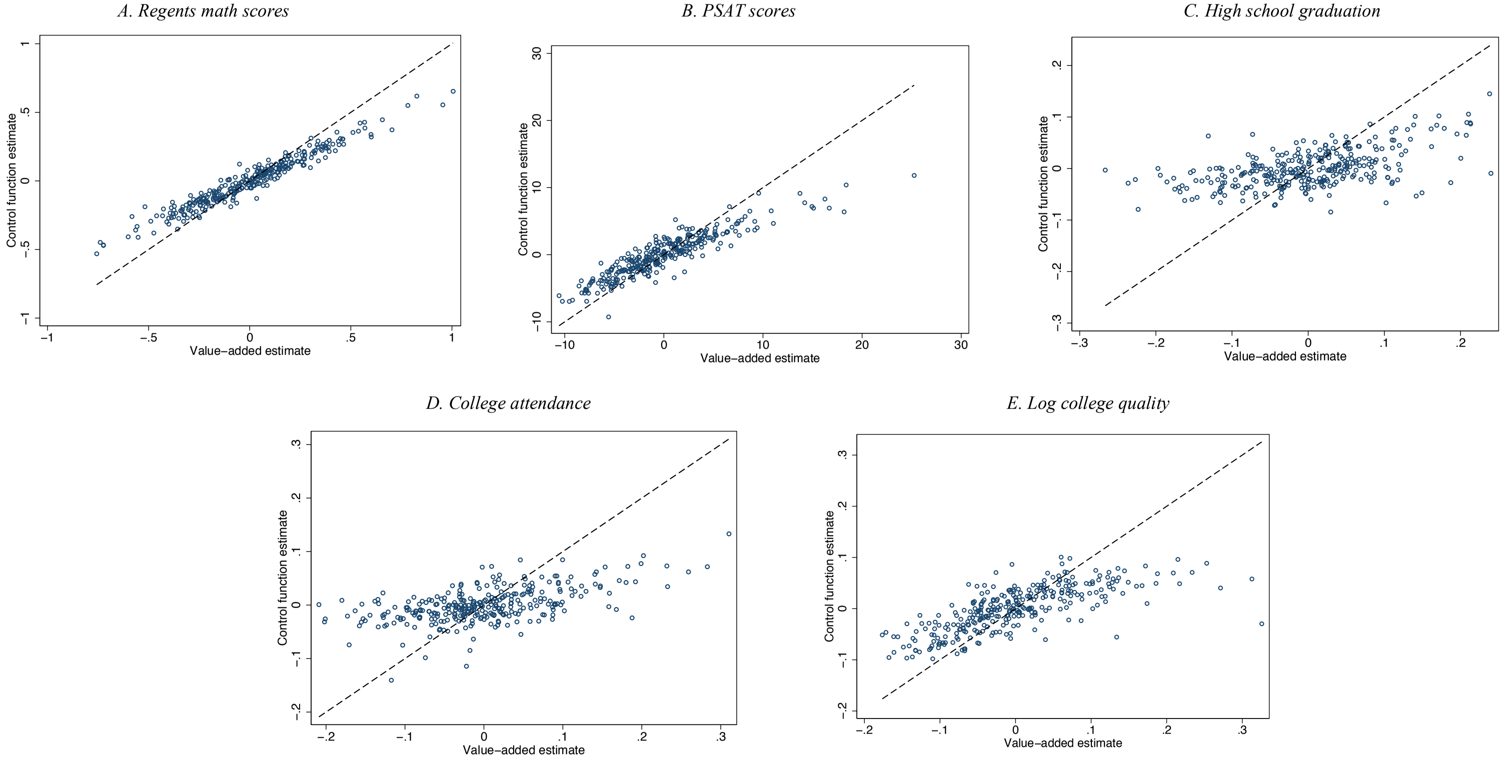
A central motivation for school choice programs is that parents' choices generate demand-side pressure for improved school productivity. We investigate this possibility by comparing estimates of school popularity and treatment effects based on rank-ordered preference data for applicants to public high schools in New York City. Parents prefer schools that enroll higher-achieving peers. Conditional on peer quality, however, parents' choices are unrelated to causal school effectiveness. Moreoever, no subgroup of parents systematically responds to causal school effectiveness. We also find no relationship between preferences for schools and estimated match quality. This indicates that choice does not lead students to sort into schools on the basis of comparative advantage in academic achievement.

This pattern of findings has important implications for the expected effects of school choice programs. Our results on match quality suggest choice is unlikely to increase allocative efficiency. Our findings regarding peer quality and average treatment effects suggest choice may create incentives for increased screening rather than academic effectiveness. If parents respond to peer quality but not causal effects, a school's easiest path to boosting its popularity is to improve the average ability of its student population. Since peer quality is a fixed resource, this creates the potential for socially costly zero-sum competition as schools invest in mechanisms to attract the best students. MacLeod and Urquiola (2015) argue that restricting a school's ability to select pupils may promote efficiency when student choices are based on school reputation. The impact of school choice on effort devoted to screening is an important empirical question for future research.

While we have shown that parents do not choose schools based on causal effects for a variety of educational outcomes, we cannot rule out the possibility that preferences are determined by effects on unmeasured outcomes. Parents may be sensitive to school safety or other non-academic amenities, for example. Our analysis also does not address why parents put more weight on peer quality than on treatment effects. If parents rely on student composition as a proxy for effectiveness, coupling school choice with credible information on causal effects may strengthen incentives for improved productivity and weaken the association between preferences and peer ability. Distinguishing between true tastes for peer quality and information frictions is another challenge for future work.

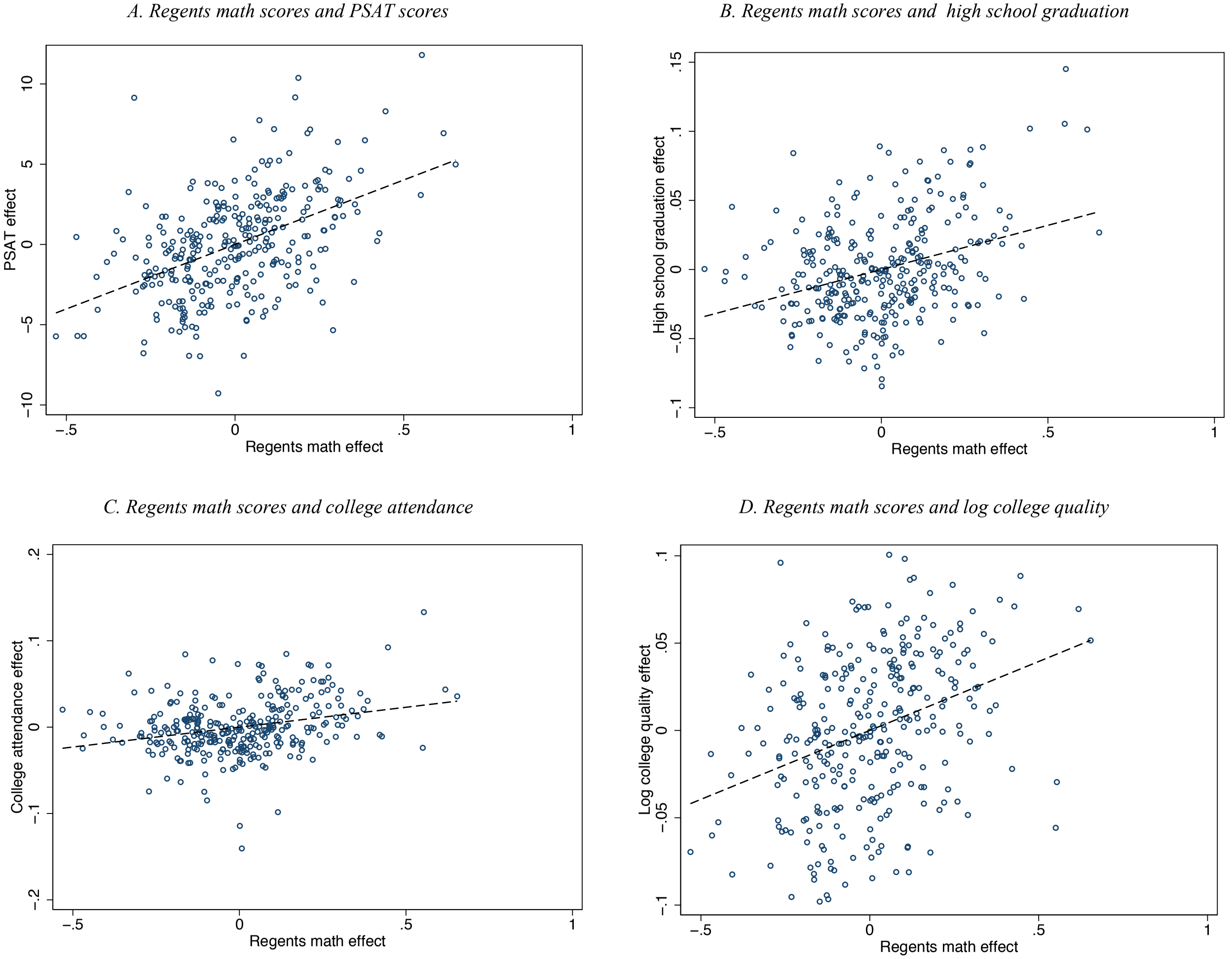
3.9 Figures

Figure 3.1: Comparison of Value-Added and Control Function Estimates of School Average Treatment Effects



Notes: This figure plots school average treatment effect (ATE) estimates from value-added models against corresponding estimates from models including control functions that adjust for selection on unobservables. Value-added estimates come from regressions of outcomes on school indicators interacted with gender, race, subsidized lunch status, the log of census tract median income, and eighth grade math and reading scores. Control function models add distance to school and predicted unobserved tastes from the choice model. Points in the figure are empirical Bayes posterior means from models fit to the distribution of school-specific estimates. Dashed lines show the 45-degree line.

Figure 3.2: Relationships between effects on test scores and effects on long run outcomes


A. Regents math scores and PSAT scores

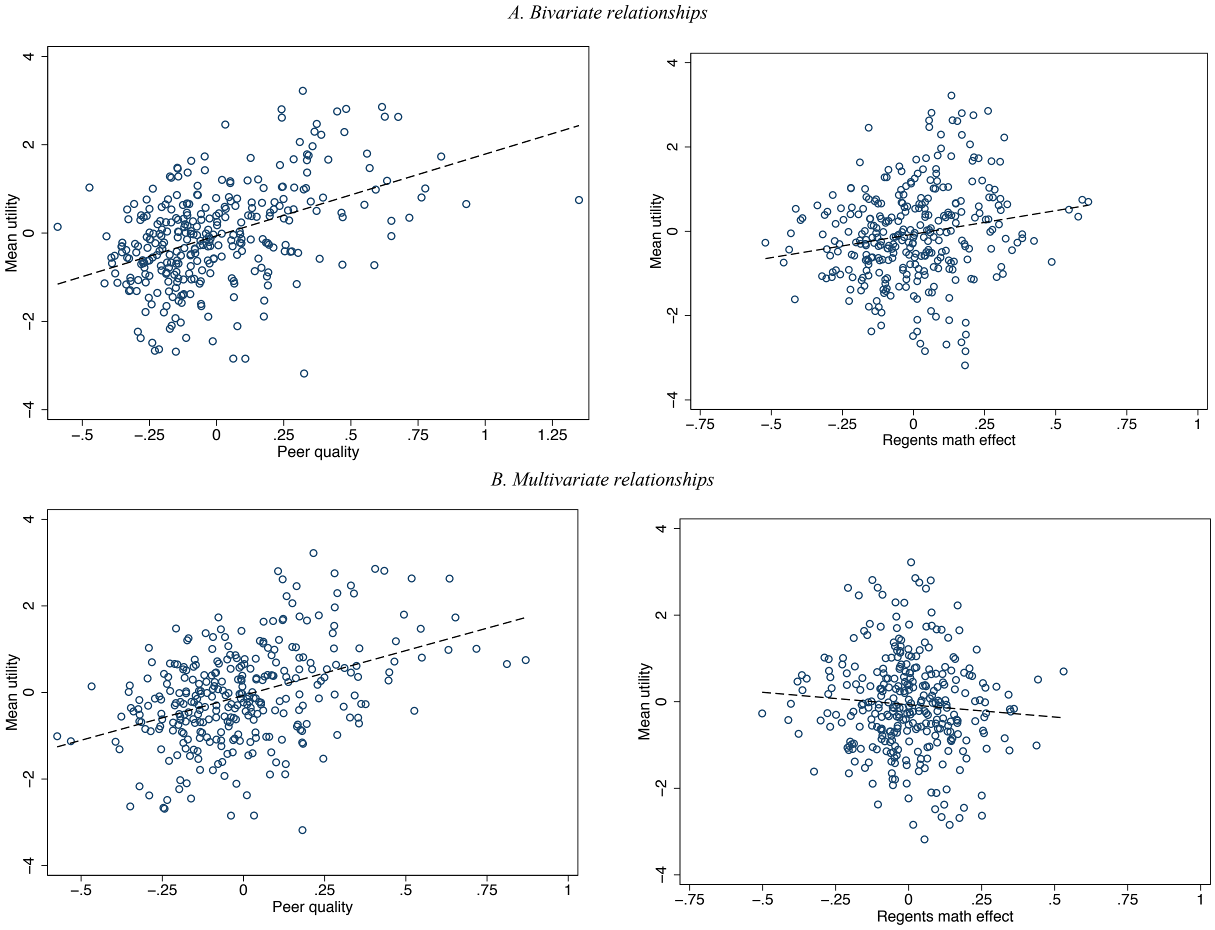
B. Regents math scores and high school graduation

C. Regents math scores and college attendance

D. Regents math scores and log college quality

Notes: This figure plots estimates of causal effects on Regents math scores against estimates of effects on longer-run outcomes. Treatment effects are empirical Bayes posterior mean estimates of school average treatment effects from control function models. Panel A plots the relationship between Regents math effects and effects on PSAT scores. Panels B, C, and D show corresponding results for high school graduation, college attendance, and log college quality.

Figure 3.3: Relationships among preferences, peer quality, and Regents math effects



A. Bivariate relationships

B. Multivariate relationships

Notes: This figure plots school mean utility estimates against estimates of peer quality and Regents math average treatment effects. Mean utilities are school average residuals from a regression of school-by-covariate cell mean utility estimates on cell indicators. Peer quality is defined as the average predicted Regents math score for enrolled students. Regents math effects are empirical Bayes posterior mean estimates of school average treatment effects from control function models. The left plot in Panel A displays the bivariate relationship between mean utility and per quality, while the right plot shows the bivariate relationship between mean utility and Regents math effects. The left plot in Panel B displays the relationship between mean utility and residuals from a regression of peer quality on Regents math effects, while the right plot shows the relationship between mean utility and residuals from a regression of Regents math effects on peer quality. Dashed lines are ordinary least squares regression lines.

3.10 Tables

Table 3.1: Descriptive Statistics for New York City Eighth Graders

		Outcome samples			
	Choice sample	Regents math	PSAT	HS graduation	College
	(1)	(2)	(3)	(4)	(5)
Female	0.497	0.518	0.532	0.500	0.500
Black	0.353	0.377	0.359	0.376	0.372
Hispanic	0.381	0.388	0.384	0.399	0.403
Subsidized lunch	0.654	0.674	0.667	0.680	0.700
Census tract median income	$50,136	$50,004	$49,993	$49,318	$49,243
Bronx	0.231	0.221	0.226	0.236	0.239
Brooklyn	0.327	0.317	0.335	0.339	0.333
Manhattan	0.118	0.118	0.119	0.116	0.116
Queens	0.259	0.281	0.255	0.250	0.253
Staten Island	0.065	0.063	0.064	0.059	0.059
Regents math score	0.000	-0.068	0.044	-0.068	-0.044
PSAT score	120	116	116	116	115
High school graduation	0.587	0.763	0.789	0.610	0.624
Attended college	0.463	0.588	0.616	0.478	0.478
College quality	$31,974	$33,934	$35,010	$31,454	$31,454
N	270157	155850	149365	230087	173254

Notes: This table shows descriptive statistics for applicants to New York City public high schools between the 2003-2004 and 2006-2007 school years. Column (1) reports average characteristics and outcomes for all applicants with complete information on preferences, demographics, and eighth-grade test scores. Columns (2)-(5) display characteristics for the Regents math, PSAT, high school graduation, and college outcome samples. Outcome samples are restricted to students with data on the relevant outcome, enrolled in for ninth grade at schools with at least 50 students for each outcome. Regents math scores are normalized to mean zero and standard deviation one in the choice sample. High school graduation equals one for students who graduate from a New York City high school within five years of the end of their eighth grade year. College attendance equals one for students enrolled in any college within two years of projected high school graduation. College quality is the mean 2014 income for individuals in the 1980-1982 birth cohorts who attended a student's college. This variable equals the mean income in the non-college population for students who did not attend college. The college outcome sample excludes students in the 2003-2004 cohort. Census tract median income is median household income measured in 2015 dollars using data from the 2011-2015 American Community Surveys. Regents math, PSAT, graduation, and college outcome statistics exclude students with missing values.

Table 3.2: Correlates of Preference Rankings for New York City High Schools

	Fraction reporting (1)	Same borough (2)	Distance (3)	Regents math score (4)
Choice 1	1.000	0.849	2.71	0.200
Choice 2	0.929	0.844	2.94	0.149
Choice 3	0.885	0.839	3.04	0.116
Choice 4	0.825	0.828	3.12	0.085
Choice 5	0.754	0.816	3.18	0.057
Choice 6	0.676	0.803	3.23	0.030
Choice 7	0.594	0.791	3.28	0.009
Choice 8	0.523	0.780	3.29	-0.013
Choice 9	0.458	0.775	3.31	-0.031
Choice 10	0.402	0.773	3.32	-0.051
Choice 11	0.345	0.774	3.26	-0.071
Choice 12	0.278	0.787	3.04	-0.107

Notes: This table reports average characteristics of New York City high schools by student preference rank. Column (1) displays fractions of student applications listing each choice. Column (2) reports the fraction of listed schools located in the same borough as a student's home address. Column (3) reports the mean distance between a student's home address and each ranked school, measured in miles. This column excludes schools outside the home borough. Column (4) shows average Regents math scores in standard deviation units relative to the New York City average.

Table 3.3: Variation in Student Preference Parameters

		Standard deviations		
	Mean	Within cells	Between cells	Total
	(1)	(2)	(3)	(4)
School mean utility	-	1.117	0.500	1.223
		(0.045)	(0.003)	(0.018)
Distance cost	0.330	-	0.120	0.120
	(0.006)		(0.005)	(0.005)
Number of students		270157		
Number of schools		316		
Number of covariate cells		360		

Notes: This table summarizes variation in school value-added and utility parameters across schools and covariate cells. Utility estimates come from rank-ordered logit models fit to student preference rankings. These models include school indicators and distance to school and are estimated separately in covariate cells defined by borough, gender, race, subsidized lunch status, an indicator for above or below the median of census tract median income, and tercile of the average of eighth grade math and reading scores. Column (1) shows the mean of the distance coefficient across cells weighted by cell size. Column (2) shows the standard deviation of school mean utilities across schools within a cell, and column (3) shows the standard deviation of a given school's mean utility across cells. School mean utilities are deviated from cell averages to account for differences in the reference category across cells. Estimated standard deviations are adjusted for sampling error by subtracting the average squared standard error of the parameter estimates from the total variance.

Table 3.4: Distributions of Peer Quality and Treatment Effect Parameters for Regents Math Scores

	Value-added model		Control function model	
	Mean	Std. dev.	Mean	Std. dev.
	(1)	(2)	(3)	(4)
Peer quality	0	0.288	0	0.305
	-	(0.012)	-	(0.012)
ATE	0	0.290	0	0.233
	-	(0.012)	-	(0.014)
Female	-0.048	0.062	-0.029	0.062
	(0.005)	(0.006)	(0.005)	(0.006)
Black	-0.112	0.130	-0.108	0.120
	(0.011)	(0.011)	(0.010)	(0.011)
Hispanic	-0.097	0.114	-0.085	0.105
	(0.010)	(0.011)	(0.010)	(0.012)
Subsidized lunch	0.001	0.052	0.026	0.054
	(0.005)	(0.006)	(0.005)	(0.006)
Log census tract median income	0.020	0.037	0.013	0.045
	(0.005)	(0.007)	(0.005)	(0.006)
Eighth grade math score	0.622	0.105	0.599	0.105
	(0.007)	(0.006)	(0.007)	(0.006)
Eighth grade reading score	0.159	0.048	0.143	0.052
	(0.004)	(0.004)	(0.004)	(0.004)
Preference coefficient (ψ_j)	-	-	-0.001	0.007
			(0.001)	(0.000)
Match coefficient (φ)	-	-	0.006	-
			(0.001)	

Notes: This table reports estimated means and standard deviations of peer quality and school treatment effect parameters for Regents math scores. Peer quality is a school's average predicted test score given the characteristics of its students. The ATE is a school's average treatment effect, and other treatment effect parameters are school-specific interactions with student characteristics. Estimates come from maximum likelihood models fit to school-specific regression coefficients. Columns (1) and (2) report estimates from an OLS regression that includes interactions of school indicators with sex, race, subsidized lunch, the log of the median income in a student's census tract, and eighth grade reading and math scores. This model also includes main effects of borough. Columns (3) and (4) show estimates from a control function model that adds distance to each school and predicted unobserved preferences from the choice model. Control functions and distance variables are set to zero for out-of-borough schools and indicators for missing values are included.

Table 3.5: Correlations of Peer Quality and Treatment Effect Parameters for Regents Math Scores

	Peer	Control function parameters							
	quality	ATE	Female	Black	Hispanic	Sub. lunch	Log tract inc.	Math score	Reading score
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ATE	0.588								
	(0.052)								
Female	0.078	0.299							
	(0.078)	(0.101)							
Black	0.006	0.107	-0.177						
	(0.077)	(0.106)	(0.142)						
Hispanic	-0.013	0.115	-0.235	0.922					
	(0.080)	(0.112)	(0.150)	(0.028)					
Subsidized lunch	0.045	-0.168	0.066	-0.038	0.004				
	(0.086)	(0.117)	(0.140)	(0.153)	(0.159)				
Log census tract income	0.035	0.068	-0.010	-0.239	-0.045	-0.280			
	(0.099)	(0.134)	(0.162)	(0.176)	(0.188)	(0.183)			
Eighth grade math score	-0.075	0.037	-0.074	-0.005	-0.007	0.060	0.027		
	(0.064)	(0.083)	(0.099)	(0.102)	(0.109)	(0.113)	(0.130)		
Eighth grade reading score	-0.418	-0.452	-0.193	-0.090	-0.078	0.004	0.086	0.256	
	(0.068)	(0.094)	(0.117)	(0.130)	(0.138)	(0.135)	(0.155)	(0.099)	
Preference coefficient (ψ_j)	0.429	0.247	0.212	-0.083	-0.058	-0.127	0.316	-0.241	-0.281
	(0.063)	(0.092)	(0.104)	(0.106)	(0.111)	(0.116)	(0.130)	(0.083)	(0.099)

Notes: This table reports estimated correlations between peer quality and school treatment effect parameters for Regents math scores. The ATE is a school's average treatment effect, and other treatment effect parameters are school-specific interactions with student characteristics. Estimates come from maximum likelihood models fit to school-specific regression coefficients from a control function model controlling for observed characteristics, distance to school and unobserved tastes from the choice model.

Table 3.6: Decomposition of School Average Outcomes

	Regents math	PSAT score/10	High school graduation	College attendance	Log college quality
	(1)	(2)	(3)	(4)	(5)
Total variance of average outcome	0.191	1.586	0.012	0.016	0.021
Variance of peer quality	0.093	0.781	0.010	0.010	0.009
Variance of ATE	0.054	0.160	0.002	0.003	0.004
Variance of match	0.008	0.027	0.002	0.002	0.001
2Cov(peer quality, ATE)	0.081	0.745	0.005	0.008	0.011
2Cov(peer quality, match)	-0.023	-0.061	-0.003	-0.003	-0.002
2Cov(ATE, match)	-0.022	-0.068	-0.004	-0.005	-0.003

Notes: This table decomposes variation in average outcomes across schools into components explained by student characteristics, school average treatment effects (ATE), and the match between student characteristics and school effects. Estimates come from control function models adjusting for selection on unobservables. Column (1) shows results for Regents math scores in standard deviation units, column (2) reports estimates for PSAT scores, column (3) displays estimates for high school graduation, column (4) reports results for college attendance, and column (5) shows results for log college quality. The first row reports the total variance of average outcomes across schools. The second row reports the variance of peer quality, defined as the average predicted outcome as a function of student characteristics and unobserved tastes. The third row reports the variance of ATE, and the fourth row displays the variance of the match effect. The remaining rows show covariances of these components.

Table 3.7: Preferences for Peer Quality and Regents Math Effects

	Value-added models				Control function models			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Peer quality	0.416		0.438	0.406	0.407		0.439	0.437
	(0.061)		(0.063)	(0.067)	(0.057)		(0.059)	(0.059)
ATE		0.244	-0.033	-0.022		0.219	-0.051	-0.047
		(0.047)	(0.046)	(0.047)		(0.046)	(0.043)	(0.043)
Match effect				-0.072				-0.172
				(0.047)				(0.054)
N	21684							

Notes: This table reports estimates from regressions of school popularity on peer quality and school effectiveness. School popularity is measured as the estimated mean utility for each school and covariate cell in the choice model from Table 4. Covariate cells are defined by borough, gender, race, subsidized lunch status, an indicator for students above the median of census tract median income, and tercile of the average of eighth grade math and reading scores. Peer quality is constructed as the average predicted Regents math score for enrolled students. Treatment effect estimates are empirical Bayes posterior mean predictions of Regents math effects. Mean utilities, peer quality, and treatment effects are scaled in standard deviation units. Columns (1)-(4) report results from value-added models, while columns (5)-(8) report results from control function models. All regressions include cell indicators and weight by the inverse of the squared standard error of the mean utility estimates. Standard errors are double- clustered by school and covariate cell.

Table 3.8: Preferences for Peer Quality and School Effectiveness by Outcome

	PSAT score		High school graduation		College attendance		Log college quality	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Peer quality		0.467		0.430		0.235		0.322
		(0.070)		(0.070)		(0.054)		(0.065)
ATE	0.325	-0.092	0.103	-0.174	0.273	0.132	0.199	0.029
	(0.056)	(0.074)	(0.045)	(0.054)	(0.048)	(0.054)	(0.059)	(0.080)
Match effect		-0.049		-0.065		-0.017		0.053
		(0.047)		(0.044)		(0.050)		(0.061)
N	21684							

Notes: This table reports estimates from regressions of school popularity on peer quality and school effectiveness separately by outcome. School popularity is measured as the estimated mean utility for each school and covariate cell in the choice model from Table 4. Covariate cells are defined by borough, gender, race, subsidized lunch status, an indicator for students above the median of census tract median income, and tercile of the average of eighth grade math and reading scores. Peer quality is constructed as the average predicted outcome for enrolled students. Treatment effect estimates are empirical Bayes posterior mean predictions from control function models. Mean utilities, peer quality, and treatment effects are scaled in standard deviation units. All regressions include cell indicators and weight by the inverse of the squared standard error of the mean utility estimates. Standard errors are double-clustered by school and covariate cell.

Table 3.9: Heterogeneity in Preferences for Peer Quality and Regents Math Effects

	By sex		By race			By subsidized lunch		By eighth grade test score tercile		
	Male	Female	Black	Hispanic	Other	Eligible	Ineligible	Lowest	Middle	Highest
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Peer quality	0.432	0.441	0.396	0.370	0.705	0.410	0.501	0.251	0.395	0.686
	(0.060)	(0.064)	(0.060)	(0.063)	(0.128)	(0.057)	(0.077)	(0.055)	(0.062)	(0.092)
ATE	-0.075	-0.021	-0.047	-0.011	-0.192	-0.036	-0.076	-0.015	-0.029	-0.117
	(0.047)	(0.043)	(0.045)	(0.044)	(0.094)	(0.042)	(0.050)	(0.042)	(0.042)	(0.059)
Match effect	-0.177	-0.169	-0.200	-0.144	-0.149	-0.180	-0.155	-0.166	-0.169	-0.125
	(0.054)	(0.054)	(0.056)	(0.066)	(0.061)	(0.054)	(0.054)	(0.061)	(0.058)	(0.055)
N	10795	10889	7467	7433	6784	11043	10641	7264	7286	7134

Notes: This table reports estimates from regressions of school popularity on peer quality and school effectiveness separately by student subgroup. School popularity is measured as the estimated mean utility for each school and covariate cell in the choice model from Table 4. Peer quality is constructed as the average predicted Regents math score for enrolled students. Treatment effect estimates are empirical Bayes posterior mean predictions of Regents math effects from control function models. Mean utilities, peer quality, and treatment effects are scaled in standard deviation units. Peer quality is constructed as the average predicted Regents math score for enrolled students. All regressions include cell indicators and weight by the inverse of the squared standard error of the mean utility estimates. Standard errors are double-clustered by school and covariate cell.

Table 3.10: Potential Achievement Gains from Ranking Schools by Effectiveness

	Observed rankings			Rankings based on effectiveness		
	Peer quality	ATE	Match	Peer quality	ATE	Match
	(1)	(2)	(3)	(4)	(5)	(6)
Choice 1	0.112	0.071	-0.037	0.286	0.427	0.162
Choice 2	0.057	0.055	-0.020	0.182	0.352	0.108
Choice 3	0.021	0.045	-0.012	0.087	0.275	0.113
Choice 4	-0.013	0.036	-0.006	0.105	0.247	0.103
Choice 5	-0.046	0.027	-0.002	0.124	0.228	0.092
Choice 6	-0.074	0.019	-0.001	0.103	0.209	0.085
Choice 7	-0.097	0.014	0.001	0.118	0.197	0.075
Choice 8	-0.114	0.012	0.001	0.099	0.169	0.066
Choice 9	-0.127	0.007	0.001	0.064	0.333	0.111
Choice 10	-0.139	0.004	0.003	0.046	0.165	0.063
Choice 11	-0.146	0.003	0.003	0.028	0.157	0.056
Choice 12	-0.156	-0.002	0.002	0.013	0.146	0.053

Notes: This table summarizes Regents math score gains that parents could achieve by ranking schools based on effectiveness. Columns (1)-(3) report average peer quality, average treatment effects, and average match quality for students' observed preference rankings. Columns (4)-(6) display corresponding statistics for hypothetical rankings that list schools in order of their treatment effects. Treatment effect estimates come from control function models. All calculations are restricted to ranked schools within the home borough.

3.11 Appendix: Data

The data used for this project were provided by the NYC Department of Education (DOE). This Appendix describes the DOE data files and explains the process used to construct our working extract from these files.

Application Data

Data on NYC high school applications are controlled by the Student Enrollment Office. We received all applications for the 2003-2004 through 2006-2007 school years. Application records include students' rank-ordered lists of academic programs submitted in each round of the application process, along with school priorities and student attributes such as special education status, race, gender, and address. The raw application files contained all applications, including private school students and first-time ninth graders who wished to change schools as well as new high school applicants. From these records we selected the set of eighth graders who were enrolled as NYC public school students in the previous school year.

Enrollment Data

We received registration and enrollment files from the Office of School Performance and Accountability (OSPA). These data include every student's grade and building code, or school ID, as of October of each school year. A separate OSPA file contains biographical information, including many of the same demographic variables from the application data. We measure demographics from the application records for variables that appeared in both files and use the OSPA file to gather additional background information such as subsidized lunch status.

OSPA also provided an attendance file with days attended and absent for each student at every school he or she attended in a given year. We use these attendance records to assign students to ninth-grade schools. If a student was enrolled in multiple schools, we use the school with the greatest number of days attended in the year following their final application to high school. A final OSPA file included scores on New York State Education Department eighth grade achievement tests. We use these test scores to assign baseline math and English Language Arts (reading) scores. Baseline scores are normalized to have mean zero and standard deviation one in our applicant sample.

Outcome Data

Our analysis studies five outcomes: Regents math scores, PSAT scores, high school graduation, college attendance, and college quality. We next describe the construction of each of these outcomes.

The Regents math test is one of five tests NYC students must pass to receive a Regents high school diploma from the state of New York. We received records of scores on all Regents tests taken between 2004 and 2008. We measured Regents math scores based on the lowest level math test offered in each year, which changed over the course of our sample. For the first three cohorts the lowest level math test offered was the Math A (Elementary Algebra and Planar Geometry) test. In 2007, the Board of Regents began administering the Math E (Integrated Algebra I) exam in addition to the Math A exam; the latter was phased out completely by 2009. We assign the earliest high school score on either of these two exams as the Regents math outcome for students in our sample. The majority of students took Math A in tenth grade, while most students taking Math E did so in ninth grade.

PSAT scores were provided to the NYC DOE by the College Board for 2003-2012. We retain PSAT scores that include all three test sections: math, reading, and writing (some subtests are missing for some observations, particularly in earlier years of our sample). If students took the PSAT multiple times, we use the score from the first attempt.

High school graduation is measured from graduation files reporting discharge status for all public school students between 2005 and 2012. These files indicate the last school attended by each student and the reason for discharge, including graduation, equivalent achievement (e.g. receiving a general equivalency diploma), or dropout. Discharge status is reported in years 4, 5, and 6 from expected graduation based on a student's year of ninth grade enrollment; our data window ends in 2012, so we only observe 4-year and 5-year high school discharge outcomes for students enrolled in eighth grade for the 2006-2007 year. We therefore focus on 5-year graduation for all four cohorts. Our graduation outcome equals one if a student received either a local diploma, a Regents diploma, or an Advanced Regents diploma within 5 years of her expected graduation date. Students not present in the graduation files are coded as not graduating.

College outcomes are measured from National Student Clearinghouse (NSC) files. The NSC records enrollment for the vast majority of post-secondary institutions, though a few important New York City-area institutions, including Rutgers and Columbia University, were not included in the NSC during our sample period.[10] The NYC DOE submitted identifying information for all NYC students graduating between 2009 and 2012 for matching to the NSC. Since many students in the 2003-04 eighth grade cohort graduated in 2008, NSC data are missing for a large fraction of this cohort. Our college outcomes are therefore defined only for the last three cohorts in the sample. For these years we code a student as attending college if she enrolled in a post-secondary institution within five years of applying to high school. This captures students who graduated from high school on time and enrolled in college the following fall, as well as students that delayed high school graduation or college enrollment by one year.

We measure college quality based on the mean 2014 incomes of students enrolled in each institution among those born between 1980 and 1982. These average incomes are reported by Chetty et al. (2017b). Fewer than 100 observations in the NSC sample failed to match to institutions in the Chetty et al. (2017b) sample. For students who enrolled in multiple postsecondary institutions, we assign the quality of the first institution attended. If

[10] In addition, about 100 parents opted out of the NSC in 2011 and 2012.

a student enrolled in multiple schools simultaneously, we use the institution with the highest mean earnings.

Matching Data Files

To construct our final analysis sample, we begin with the set of high school applications submitted by students enrolled in eighth grade between the 2003-2004 and 2006-2007 school years. We match these applications to the student enrollment file using a unique student identifier known as the OSISID and retain individuals that appear as eighth graders in both data sets. If a student submits multiple high school applications as an eighth grader, we select the final application for which data is available. We then use the OSISID to match applicant records to the OSPA attendance and test scores files (used to assign ninth grade enrollment and baseline test scores), and the Regents, PSAT, graduation, and NSC outcome files.

This merged sample is used to construct the set of 316 high schools that enrolled at least 50 students with observations for each of the five outcomes, excluding selective schools that do not participate in the main DA round. The final choice sample includes the set of high school applicants reporting at least one of these 316 schools on their preference lists. The five outcome samples are subsets of the choice sample with observed data on the relevant outcome and enrolled in one of our sample high schools for ninth grade. Table 3.11 displays the impact of each restriction on sample size for the four cohorts in our analysis sample.

3.12 Appendix: Econometric Methods

Rank-Ordered Control Functions

This section provides formulas for the rank-ordered control functions in equation (3.8). The choice model is

$$U_{ij} = \delta_{c(X_i)j} - \tau_{c(X_i)} D_{ij} + \eta_{ij} = V_{ij} + \eta_{ij},$$

where $V_{ij} \equiv \delta_{c(X_i)j} - \tau_{c(X_i)} D_{ij}$ represents the observed component of student i's utility for school j and η_{ij} is the unobserved component. The control functions are given by $\lambda_{ij} = E[\eta_{ij} - \mu_\eta | X_i, D_i, R_i] = E[\eta_{ij} | R_i, V_i] - \mu_\eta$, where $V_i = (V_{i1}, ..., V_{iJ})'$. To compute the conditional mean of η_{ij}, it will be useful to define the following functions for any set of mean utilities S and subset $S' \subseteq S$:

$$P\left(S'|S\right) = \frac{\sum\limits_{v \in S'} \exp(v)}{\sum\limits_{v \in S} \exp(v)},$$

$$\mathcal{I}(S) = \mu_\eta + \log\left(\sum\nolimits_{v \in S} \exp(v)\right).$$

$P(S'|S)$ gives the probability that an individual chooses an option in S' from the set S when the value of each option is the sum of its mean utility and an extreme value type I error term, while $\mathcal{I}(S)$ gives the expected maximum utility of choosing an option in S, also known as the inclusive value. We provide expressions for the control functions for two cases: (1) when a student ranks all available alternatives, and (2) when the student leaves some alternatives unranked.

All alternatives ranked

Control function for the highest-ranked alternative

Without loss of generality, label alternatives in decreasing order of student i's preferences, so that $R_{ij} = j$ for $j = 1...J$. The control function associated with the highest ranked alternative is

$$\begin{aligned}\lambda_{i1} &= -(V_{i1} + \mu_\eta) + E\left[U_{i1}|R_i, V_i\right] \\ &= -(V_{i1} + \mu_\eta) + \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{u_1}\int_{-\infty}^{u_2}\cdots\int_{-\infty}^{u_{J-1}}\left[u_1\prod_{k=1}^{J} f(u_k|V_{ik})\right]du_J...du_2du_1}{\prod_{k=1}^{J-1} P\left(V_{ik}|S_i(k)\right)}\end{aligned}$$

where $S_i(m) = \{V_{ik} : k \geq m\}$ and $f(u|V) = \exp\left(V - u - \exp(V - u)\right)$ is the density function of a Gumbel random variable with location parameter V. This simplifies to

$$\begin{aligned}\lambda_{i1} &= -(V_{i1} + \mu_\eta) + \frac{\prod_{k=1}^{J} P\left(V_{ik}|S_i(k)\right) \times \mathcal{I}(S_i(1))}{\prod_{k=1}^{J-1} P\left(V_{ik}|S_i(k)\right)} \\ &= -V_{i1} + \left(\mathcal{I}(S_i(1)) - \mu_\eta\right) \\ &= -\log\left(P(V_{i1}|S_i(1))\right)\end{aligned}$$

which coincides with the control function for the most preferred alternative in the multinomial logit model of Dubin and McFadden (1984). This shows that knowledge of the rankings of less-preferred alternatives does not affect the expected utility associated with the best choice.

Control functions for lower-ranked alternatives

To work out λ_{ij} for $j > 1$, define the following functions:

$$\begin{aligned}G_{i0}(u) &= 1 \\ G_{ij}(u) &= \int_{u}^{\infty} f(x|V_{ij})G_{i(j-1)}(x)dx,\ j = 1...J\end{aligned}$$

It can be shown that[11]

$$G_{ij}(u)=\sum_{m=1}^{j}(-1)^{j+m}\left[\prod_{q=1}^{m-1}P\left(V_{iq}|V_{iq}...V_{i(m-1)}\right)\right]\left[\prod_{n=m}^{j}P(V_{in}|V_{im}...V_{in})\right][1-F(u|\mathcal{I}(V_m...V_j)-\mu_\eta)] \tag{3.13}$$

where $F(u|V)=\exp(-\exp(V-u))$ is the Gumbel CDF with location V. The first product is defined as 1 when $m=1$. Then for $j>1$, we have

$$\begin{aligned}
\lambda_{ij} &= -(V_{ij}+\mu_\eta) \\
&\quad + \frac{\int_{-\infty}^{\infty}\int_{u_j}^{\infty}\int_{u_{j-1}}^{\infty}...\int_{u_2}^{\infty}\int_{-\infty}^{u_j}\int_{-\infty}^{u_{j+1}}...\int_{-\infty}^{u_{J-1}}\left[u_j\prod_{k=1}^{J}f(u_k|V_{ik})\right]du_J...du_{j+1}du_1...du_j}{\prod_{k=1}^{J-1}P\left(V_{ik}|S_i(k)\right)} \\
&= -(V_{ij}+\mu_\eta)+\frac{\int_{-\infty}^{\infty}u_jf(u_j|\mathcal{I}(S_i(j))-\mu_\eta)G_{i(j-1)}(u_j)du_j}{\prod_{k=1}^{j-1}P\left(V_{ik}|S_i(k)\right)} \\
&= -\log\left(P\left(V_{ij}|S_i(j)\right)\right) \\
&\quad + \sum_{m=1}^{j-1}(-1)^{(j-1)+m}\left(\frac{\prod_{n=m}^{j-1}\left(\frac{1-P(V_{im}...V_{in}|S_i(m))}{P(V_{im}...V_{in}|S_i(m))}\right)}{\prod_{q=1}^{m-1}P\left(V_{iq}...V_{i(m-1)}|S_i(q)\right)}\right)\log\left(1-P\left(V_{im}...V_{i(j-1)}|S_i(m)\right)\right)
\end{aligned}$$

The expected utility of j depends on both the probability of choosing j over all lower-ranked alternatives and the choice probabilities of alternatives preferred to j. Again, as shown previously for $j=1$, the expected utility for any ranked option does not depend on the rank order of less-preferred alternatives.

Unranked alternatives

To derive the control functions for a case in which some alternatives are unranked, assign arbitrary labels $\ell(i)+1....J$ to unranked schools. The control functions for all ranked alternatives can be obtained by defining a composite unranked alternative with observed utility $V_{iu}=\mathcal{I}\left(V_{ik}:k>\ell(i)\right)-\mu_\eta$ and treating this as the lowest-ranked option. The control

[11] The derivation of this non-recursive expression requires the following identity, which holds $\forall j\geq k\geq 1$, $a_k>0$:

$$\sum_{m=1}^{j}(-1)^{k+j}\left[\prod_{p=1}^{m-1}\left(\frac{1}{\sum\limits_{m=p}^{m-1}a_m}\right)\right]\left[\prod_{q=m}^{j}\left(\frac{1}{\sum\limits_{n=m}^{q}a_n}\right)\right]=\left[\prod_{s=1}^{j}\left(\frac{1}{\sum\limits_{r=s}^{j}a_r}\right)\right] \tag{3.12}$$

with the first product defined as 1 when $m=1$. We prove this identity in Appendix 3.13.

function for an unranked alternative j is defined by the expression

$$\begin{aligned}
\lambda_{ij} + (V_{ij} + \mu_\eta) &= E\left[U_{ij} | U_{i1} > ... > U_{i\ell(i)},\ U_{i\ell(i)} > U_{ik}\ \forall k > \ell(i), V_i\right] \\
&= \frac{\int_{-\infty}^{\infty}\int_{u_j}^{\infty}\int_{u_{\ell(i)}}^{\infty} ... \int_{u_2}^{\infty}\int_{-\infty}^{u_{\ell(i)}} ... \int_{-\infty}^{u_{\ell(i)}} u_j \prod_{k=1}^{J} f(u_k|V_{ik}) du_{\ell(i)+1}...du_{j-1}du_{j+1}...du_J du_1...du_{\ell(i)}du_j}{\prod_{k=1}^{\ell(i)} P(V_{ik}|S_i(k))} \\
&= \frac{\int_{-\infty}^{\infty} u_j f(u_j|V_{ij}) \left[\int_{u_j}^{\infty} f\left(u_{\ell(i)} | \mathcal{I}(S_i^{-j}(\ell(i))) - \mu_\eta\right) G_{i(\ell(i)-1)}(u_{\ell(i)}) du_{\ell(i)}\right] du_j}{P(V_{i\ell(i)}|S_i^{-j}(\ell(i)))^{-1} \times \prod_{k=1}^{\ell(i)} P(V_{ik}|S_i(k))}
\end{aligned}$$

where $S_i^{-j}(m) = \{V_{ik} : k \geq m\} \backslash \{V_{ij}\}$ is the set of i's mean utilities for alternatives m and higher excluding alternative j. When $\ell(i) = 1$, we have $G_{i(\ell(i)-1)}(u_\ell) = 1$ and this expression collapses to

$$\lambda_{ij} = \frac{P(V_{ij}|S_i(1))}{1 - P(V_{ij}|S_i(1))} \log\left(P(V_{ij}|S_i(1))\right)$$

which is the expression derived by Dubin and McFadden (1984) for the expected errors of alternatives that are not selected in the multinomial logit model. For $\ell(i) > 1$, we have

$$\lambda_{ij} = \sum_{m=1}^{\ell(i)} (-1)^{m+\ell(i)} \left(\frac{\prod_{n=m}^{\ell(i)-1} \left(\frac{1-P(V_{im}...V_{in}|S_i(m))}{P(V_{im}...V_{in}|S_i(m))}\right)}{\prod_{q=1}^{m-1} P\left(V_{iq}...V_{i(m-1)}|S_i(q)\right)} \right) \left[\frac{P\left(V_{ij}|S_i(m)\right)}{1 - P\left(V_{ij}|S_i(m)\right)}\right] \log\left(P\left(V_{ij}|S_i(m)\right)\right)$$

Similar to the multinomial logit case, λ_{ij} is a depends on the choice probablity of j relative to other ranked options, and does not depend on the probability of other unranked options.

Two-Step Score Bootstrap

We use a two-step modification of the score bootstrap of Kline and Santos (2012) to conduct inference for the control function models. Let $\Delta = (\delta_{11}...\delta_{1J}, \tau_1...\delta_{C1}...\delta_{CJ}, \tau_C)'$ denote the vector of choice model parameters for all covariate cells. Maximum likelihood estimates of these parameters are given by:

$$\hat{\Delta} = \arg\max_{\Delta} \sum_i \log \mathcal{L}(R_i|X_i, D_i; \Delta),$$

where $\mathcal{L}(R_i|X_i, D_i; \Delta)$ is the likelihood function defined in Section 3.4, now explicitly written as a function of the choice model parameters.

Let $\Gamma = (\alpha_1, \beta_1', \psi_1...\alpha_J, \beta_J', \psi_J, \gamma', \varphi)'$ denote the vector of outcome equation parameters. Second-step estimates of these parameters are

$$\hat{\Gamma} = \left[\sum_i W_i(\hat{\Delta})W_i(\hat{\Delta})'\right]^{-1} \times \sum_i W_i(\hat{\Delta})Y_i,$$

where $W_i(\Delta)$ is the vector of regressors in equation (3.8). This vector depends on Δ through the control functions $\lambda_j(X_i, D_i, R_i; \Delta)$, which in turn depend on the choice model parameters as described previously.

The two-step score bootstrap adjusts inference for the extra uncertainty introduced by the first-step estimates while avoiding the need to recalculate $\hat{\Delta}$ or to analytically derive the influence of $\hat{\Delta}$ on $\hat{\Gamma}$. The first step directly applies the approach in Kline and Santos (2012) to the choice model estimates. This approach generates a bootstrap distribution for $\hat{\Delta}$ by taking repeated Newton-Raphson steps from the full-sample estimates, randomly reweighting each observation's score contribution. The bootstrap estimate of Δ in trial $b \in \{1...B\}$ is:

$$\hat{\Delta}^b = \hat{\Delta} - \left[\sum_i \left(\frac{\partial^2 \log \mathcal{L}\big(R_i|X_i, D_i; \hat{\Delta}\big)}{\partial\Delta\partial\Delta'}\right)\right]^{-1} \times \sum_i \zeta_i^b \left(\frac{\partial \log \mathcal{L}(R_i|X_i, D_i; \hat{\Delta})}{\partial\Delta}\right),$$

where the ζ_i^b are *iid* random weights satisfying $E\left[\zeta_i^b\right] = 0$ and $E\left[(\zeta_i^b)^2\right] = 1$. We draw these weights from a standard normal distribution.

Next, we use an additional set of Newton-Raphson steps to generate a bootstrap distribution for $\hat{\Gamma}$. The second-step bootstrap estimates are:

$$\hat{\Gamma}^b = \hat{\Gamma} - \left[\sum_i W_i(\hat{\Delta})W_i(\hat{\Delta})'\right]^{-1} \times \sum_i \left[-\zeta_i^b W_i(\hat{\Delta})(Y_i - W_i(\hat{\Delta})'\hat{\Gamma}) - W_i(\hat{\Delta}^b)(Y_i - W_i(\hat{\Delta}^b)'\hat{\Gamma})\right].$$

The second term in the last sum accounts for the additional variability in the second-step score due to the first-step estimate $\hat{\Delta}$. We construct standard errors and conduct hypothesis tests involving Γ using the distribution of $\hat{\Gamma}^b$ across bootstrap trials.

Empirical Bayes Shrinkage

We next describe the empirical Bayes shrinkage procecure summarized in Section 3.4. Value-added or control function estimation produces a set of school-specific parameter estimates, $\left\{\hat{\theta}_j\right\}_{j=1}^J$. Under the hierarchical model (3.10), the likelihood of the estimates for school j conditional on the latent parameters θ_j and the sampling variance matrix Ω_j is:

$$\mathcal{L}\left(\hat{\theta}_j|\theta_j, \Omega_j\right) = (2\pi)^{-T/2}\,|\Omega_j|^{-1/2} \exp\left(-\tfrac{1}{2}(\hat{\theta}_j - \theta_j)'\Omega_j^{-1}(\hat{\theta}_j - \theta_j)\right),$$

where $T = \dim(\theta_j)$. We estimate Ω_j using conventional asymptotics for the value-added models and the bootstrap procedure described in the previous section for the control function models. Our approach therefore requires school-specific samples to be large enough for these asymptotic approximations to be accurate.

An integrated likelihood function that conditions only on the hyperparameters is:

$$\begin{aligned}\mathcal{L}^I(\hat{\theta}_j|\mu_\theta,\Sigma_\theta,\Omega_j) &= \int \mathcal{L}(\hat{\theta}_j|\theta_j,\Omega_j)dF(\theta_j|\mu_\theta,\Sigma_\theta)\\ &= (2\pi)^{-T/2}\left|\Omega_j+\Sigma_\theta\right|^{-1/2}\exp\left(-\tfrac{1}{2}(\hat{\theta}_j-\mu_\theta)'\left(\Omega_j+\Sigma_\theta\right)^{-1}(\hat{\theta}_j-\mu_\theta)\right).\end{aligned}$$

EB estimates of the hyperparameters are then

$$\left(\hat{\mu}_\theta,\hat{\Sigma}_\theta\right) = \arg\max_{\mu_\theta,\Sigma_\theta}\sum_j \log\mathcal{L}^I\left(\hat{\theta}_j|\mu_\theta,\Sigma_\theta,\hat{\Omega}_j\right),$$

where $\hat{\Omega}_j$ estimates Ω_j.

By standard arguments, the posterior distribution for θ_j given the estimate $\hat{\theta}_j$ is

$$\theta_j|\hat{\theta}_j \sim N\left(\theta_j^*,\Omega_j^*\right),$$

where

$$\theta_j^* = \left(\Omega_j^{-1}+\Sigma_\theta^{-1}\right)^{-1}\left(\Omega_j^{-1}\hat{\theta}_j+\Sigma_j^{-1}\mu_\theta\right),$$

$$\Omega_j^* = \left(\Omega_j^{-1}+\Sigma_\theta^{-1}\right)^{-1}.$$

We form EB posteriors by plugging $\hat{\Omega}_j$, $\hat{\mu}_\theta$ and $\hat{\Sigma}_\theta$ into these formulas.

3.13 Appendix: Proof of Identity to Express Control Function Non-Recursively

Claim: for all $j > 1$ and positive scalars a_k with $k = 1, ..., j$, we have:

$$\sum_{k=1}^{j-1}(-1)^{k+(j-1)}\left[\prod_{\substack{z=1\\z\neq k-1}}^{j-1}\prod_{\substack{\ell=1\\\ell\neq k}}^{z}\left(\sum_{p=\ell}^{z}a_p\right)\right] = \prod_{z=1}^{j-2}\prod_{\ell=1}^{z}\left(\sum_{p=\ell}^{z}a_p\right) \tag{3.14}$$

$$\Leftrightarrow \sum_{k=1}^{j}(-1)^{k+j}\left[\prod_{\substack{z=1\\z\neq k-1}}^{j-1}\prod_{\substack{\ell=1\\\ell\neq k}}^{z}\left(\sum_{p=\ell}^{z}a_p\right)\right] = 0 \tag{3.14$'$}$$

where sums with zero terms or over undefined indices are defined as zero and products over said terms are defined as 1. The above claim is identical to the identity expressed in (3.12) from Appendix 3.12. To simplify the above notation, define the following:

$$\mu_{\ell,z} \equiv \sum_{p=\ell}^{z}a_p \tag{3.15}$$

Now, our induction claim is as follows:

$$\sum_{k=1}^{j-1}(-1)^{k+(j-1)}\left[\prod_{\substack{z=1\\z\neq k-1}}^{j-1}\prod_{\substack{\ell=1\\\ell\neq k}}^{z}\mu_{\ell,z}\right]=\prod_{z=1}^{j-2}\prod_{\ell=1}^{z}\mu_{\ell,z} \tag{3.16}$$

$$\Leftrightarrow\sum_{k=1}^{j}(-1)^{k+j}\left[\prod_{\substack{z=1\\z\neq k-1}}^{j-1}\prod_{\substack{\ell=1\\\ell\neq k}}^{z}\mu_{\ell,z}\right]=0 \tag{3.16$'$}$$

We will refer to Equation (3.16) for our claim. We will prove using induction.

- The base case $j=2$ is trivial. In (3.16), both sides of the equation have the outer products contain zero elements, which by definition is 1.
- Let's also check with $j=3$, the first "non-trivial" case.
  - $LHS=(-1)\times\mu_{2,2}+(1)\times\mu_{1,2}=a_1$
  - $RHS=\mu_{1,1}=a_1$

To finish the proof, we will show that (3.16) implies the following:

$$\sum_{k=1}^{j}(-1)^{k+j}\left[\prod_{\substack{z=1\\z\neq k-1}}^{j}\prod_{\substack{\ell=1\\\ell\neq k}}^{z}\mu_{\ell,z}\right]=\prod_{z=1}^{j-1}\prod_{\ell=1}^{z}\mu_{\ell,z} \tag{3.17}$$

Before we begin, we will make two observations that will be used in our proof:

1. By definition of $\mu_{\ell,z}$:
$$\mu_{v,r}=0\ \ \forall v>r \tag{3.18}$$
2. To prove (3.17), it suffices to show that $\Gamma^j_m=0$ for $m=0,...,j-2$, where Γ^j_m is defined as follows:

$$\Gamma^j_m=\sum_{k=1}^{j}(-1)^{k+j}\left[\prod_{\substack{z=1\\z\neq k-1}}^{j-1}\prod_{\substack{\ell=1\\\ell\neq k}}^{z}\mu_{\ell,z}\right]\left[\sum_{\substack{S\subset 2^{\{1,\ldots,j\}}\\k\notin S\\|S|=m}}\left(\prod_{x\in S}\mu_{x,j-1}\right)\right]\quad m=0,...,j-1 \tag{3.19}$$

This is because the left hand side of (3.17) can be rewritten as follows:

$$\sum_{k=1}^{j}(-1)^{k+j}\left[\prod_{\substack{z=1\\z\neq k-1}}^{j}\prod_{\substack{\ell=1\\\ell\neq k}}^{z}\mu_{\ell,z}\right]\equiv\sum_{m=0}^{j-1}\Gamma^j_m[a_j]^{j-1-m} \tag{3.20}$$

To show $\Gamma^j_m = 0$ for all $m = 0, ..., j-2$, I will first show that our inductive hypothesis implies the following for all n, where $0 \leq n \leq m < j$:

$$\Gamma^j_m = \left[\prod_{y=1}^{n}\prod_{x=1}^{j-y}\mu_{x,j-y}\right]\sum_{k=1}^{j-n}(-1)^{k+(j-n)}\left[\prod_{\substack{z=1\\ z\neq k-1}}^{j-n-1}\prod_{\substack{\ell=1\\ \ell\neq k}}^{z}\mu_{\ell,z}\right]\left[\sum_{\substack{S\subset 2^{\{1,\dots,j-n\}}\\ k\notin S\\ |S|=m-n}}\left(\prod_{x\in S}\mu_{x,j-1}\right)\right] \tag{3.21}$$

By (3.19), the definition of Γ^j_m, Equation (3.21) holds for $n = 0$. Assuming true for $n-1 \geq 0$, we have:

$$\begin{aligned}
\Gamma^j_m = & \left[\prod_{y=1}^{n-1}\prod_{x=1}^{j-y}\mu_{x,j-y}\right]\sum_{k=1}^{j-n+1}(-1)^{k+1+(j-n)}\left[\prod_{\substack{z=1\\ z\neq k-1}}^{j-n}\prod_{\substack{\ell=1\\ \ell\neq k}}^{z}\mu_{\ell,z}\right]\left[\sum_{\substack{S\subset 2^{\{1,\dots,j-n+1\}}\\ k\notin S\\ |S|=m-n+1}}\left(\prod_{x\in S}\mu_{x,j-1}\right)\right] \\
= & \left[\prod_{y=1}^{n-1}\prod_{x=1}^{j-y}\mu_{x,j-y}\right]\sum_{k=1}^{j-n}(-1)^{k+1+(j-n)}\left[\prod_{\substack{z=1\\ z\neq k-1}}^{j-n}\prod_{\substack{\ell=1\\ \ell\neq k}}^{z}\mu_{\ell,z}\right]\left[\sum_{\substack{S\subset 2^{\{1,\dots,j-n+1\}}\\ k\notin S\\ |S|=m-n+1}}\left(\prod_{x\in S}\mu_{x,j-1}\right)\right] \\
& + \left[\prod_{y=1}^{n-1}\prod_{x=1}^{j-y}\mu_{x,j-y}\right]\left[\prod_{z=1}^{j-n-1}\prod_{\ell=1}^{z}\mu_{\ell,z}\right]\left[\sum_{\substack{S\subset 2^{\{1,\dots,j-n\}}\\ |S|=m-n+1}}\left(\prod_{x\in S}\mu_{x,j-1}\right)\right] \\
= & -\left[\prod_{y=1}^{n-1}\prod_{x=1}^{j-y}\mu_{x,j-y}\right]\sum_{k=1}^{j-n}(-1)^{k+(j-n)}\left[\prod_{\substack{z=1\\ z\neq k-1}}^{j-n}\prod_{\substack{\ell=1\\ \ell\neq k}}^{z}\mu_{\ell,z}\right]\left[\sum_{\substack{S\subset 2^{\{1,\dots,j-n+1\}}\\ k\notin S\\ |S|=m-n+1}}\left(\prod_{x\in S}\mu_{x,j-1}\right)\right] \\
& + \left[\prod_{y=1}^{n-1}\prod_{x=1}^{j-y}\mu_{x,j-y}\right]\sum_{k=1}^{j-n}(-1)^{k+(j-n)}\left[\prod_{\substack{z=1\\ z\neq k-1}}^{j-n}\prod_{\substack{\ell=1\\ \ell\neq k}}^{z}\mu_{\ell,z}\right]\left[\sum_{\substack{S\subset 2^{\{1,\dots,j-n\}}\\ |S|=m-n+1}}\left(\prod_{x\in S}\mu_{x,j-1}\right)\right]
\end{aligned}$$

Continued...

$$\Gamma^j_m = \left[\prod_{y=1}^{n-1}\prod_{x=1}^{j-y}\mu_{x,j-y}\right]\sum_{k=1}^{j-n}(-1)^{k+(j-n)}\mu_{k,j-n}\left[\prod_{\substack{z=1\\z\neq k-1}}^{j-n}\prod_{\substack{\ell=1\\\ell\neq k}}^{z}\mu_{\ell,z}\right]\left[\sum_{\substack{S\subset 2^{\{1,\dots,j-n\}}\\k\notin S\\|S|=m-n}}\left(\prod_{x\in S}\mu_{x,j-1}\right)\right]$$

$$= \left[\prod_{x=1}^{j-n}\mu_{x,j-n}\right]\left[\prod_{y=1}^{n-1}\prod_{x=1}^{j-y}\mu_{x,j-y}\right]\sum_{k=1}^{j-n}(-1)^{k+j-n}\left[\prod_{\substack{z=1\\z\neq k-1}}^{j-n-1}\prod_{\substack{\ell=1\\\ell\neq k}}^{z}\mu_{\ell,z}\right]\left[\sum_{\substack{S\subset 2^{\{1,\dots,j-n\}}\\k\notin S\\|S|=m-n}}\left(\prod_{x\in S}\mu_{x,j-1}\right)\right]$$

$$= \left[\prod_{y=1}^{n}\prod_{x=1}^{j-y}\mu_{x,j-y}\right]\sum_{k=1}^{j-n}(-1)^{k+(j-n)}\left[\prod_{\substack{z=1\\z\neq k-1}}^{j-n-1}\prod_{\substack{\ell=1\\\ell\neq k}}^{z}\mu_{\ell,z}\right]\left[\sum_{\substack{S\subset 2^{\{1,\dots,j-n\}}\\k\notin S\\|S|=m-n}}\left(\prod_{x\in S}\mu_{x,j-1}\right)\right]$$

Therefore, our inductive hypothesis implies (3.21) holds. Let $n = m$ in (3.21). Then we have the following for $m < j-1$:

$$\Gamma^j_m = \left[\prod_{y=1}^{m}\prod_{x=1}^{j-y}\mu_{x,j-y}\right]\sum_{k=1}^{j-m}(-1)^{k+(j-m)}\left[\prod_{\substack{z=1\\z\neq k-1}}^{j-m-1}\prod_{\substack{\ell=1\\\ell\neq k}}^{z}\mu_{\ell,z}\right]\underbrace{\left[\sum_{\substack{S\subset 2^{\{1,\dots,j-m\}}\\k\notin S\\|S|=m-m}}\left(\prod_{x\in S}\mu_{x,j-1}\right)\right]}_{=1}$$

$$= \left[\prod_{y=1}^{m}\prod_{x=1}^{j-y}\mu_{x,j-y}\right]\underbrace{\sum_{k=1}^{j-m}(-1)^{k+(j-m)}\left[\prod_{\substack{z=1\\z\neq k-1}}^{j-m-1}\prod_{\substack{\ell=1\\\ell\neq k}}^{z}\mu_{\ell,z}\right]}_{=0\text{ (by (3.16}'\text{))}}$$

$$= 0$$

When $m = n = j - 1$:

$$\Gamma^j_{j-1} = \left[\prod_{y=1}^{(j-1)} \prod_{x=1}^{j-y} \mu_{x,j-y}\right] \sum_{k=1}^{j-(j-1)} (-1)^{k+j-(j-1)} \left[\prod_{\substack{z=1\\ z\neq k-1}}^{j-(j-1)-1} \prod_{\substack{\ell=1\\ \ell\neq k}}^{z} \mu_{\ell,z}\right] \underbrace{\left[\sum_{\substack{S\subset 2^{\{1,\ldots,j-n\}}\\ k\notin S\\ |S|=0}} \left(\prod_{x\in S} \mu_{x,j-1}\right)\right]}_{=1}$$

$$= \left[\prod_{z=1}^{j-1} \prod_{\ell=1}^{z} \mu_{\ell,z}\right] \underbrace{\left[\prod_{\substack{z=1\\ z\neq 0}}^{0} \prod_{\substack{\ell=1\\ \ell\neq 1}}^{z} \mu_{\ell,z}\right]}_{=1}$$

$$= \left[\prod_{z=1}^{j-1} \prod_{\ell=1}^{z} \mu_{\ell,z}\right]$$

Combining the above results with (3.20), we have:

$$\begin{aligned} \sum_{k=1}^{j} (-1)^{k+j} \left[\prod_{\substack{z=1\\ z\neq k-1}}^{j} \prod_{\substack{\ell=1\\ \ell\neq k}}^{z} \mu_{\ell,z}\right] &= \sum_{m=0}^{j-1} \Gamma^j_m [a_j]^{j-1-m} \\ &= \Gamma^j_{j-1} + \sum_{m=0}^{j-2} \Gamma^j_m [a_j]^{j-1-m} \\ &= \left[\prod_{z=1}^{j-1} \prod_{\ell=1}^{z} \mu_{\ell,z}\right] \end{aligned}$$

3.14 Appendix: Additional Tables

Table 3.11: Sample Restrictions

	All cohorts	2003-2004	2004-2005	2005-2006	2006-2007
	(1)	(2)	(3)	(4)	(5)
All NYC eighth graders	368,603	89,671	93,399	94,015	91,518
In public school	327,948	78,904	83,112	84,067	81,865
With baseline demographics	276,797	68,507	67,555	68,279	72,456
With address data	275,405	67,644	67,377	68,108	72,276
In preference sample	270,157	66,125	66,004	67,163	70,865
In Regents math sample	155,850	40,994	41,022	39,177	34,657
In PSAT sample	149,365	31,563	37,502	39,480	40,820
In high school graduation sample	230,087	56,833	56,979	57,803	58,472
In college sample	173,254	0	56,979	57,803	58,472

Notes: This table displays the selection criteria for inclusion in the final analysis samples. Preference models are estimated using the sample in the fourth row, and school effects are estimated using the samples in the remaining rows.

Table 3.12: Correlations of Peer Quality and Treatment Effect Parameters for Regents Math Scores, Value-Added Model

	Peer	Value-added parameters						
	quality	ATE	Female	Black	Hispanic	Sub. lunch	Log tract inc.	Math score
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ATE	0.531							
	(0.042)							
Female	0.133	0.232						
	(0.077)	(0.082)						
Black	-0.033	-0.007	-0.287					
	(0.074)	(0.082)	(0.133)					
Hispanic	-0.002	-0.028	-0.414	0.939				
	(0.077)	(0.086)	(0.135)	(0.022)				
Subsidized lunch	0.093	-0.133	0.098	-0.027	0.065			
	(0.088)	(0.097)	(0.145)	(0.151)	(0.155)			
Log census tract income	-0.288	-0.108	-0.210	-0.140	-0.048	-0.200		
	(0.111)	(0.129)	(0.185)	(0.202)	(0.212)	(0.220)		
Eighth grade math score	-0.108	0.033	-0.104	-0.005	0.054	0.012	-0.083	
	(0.064)	(0.069)	(0.098)	(0.100)	(0.105)	(0.118)	(0.150)	
Eighth grade reading score	-0.564	-0.425	-0.036	-0.065	-0.064	0.071	0.374	0.244
	(0.065)	(0.079)	(0.124)	(0.123)	(0.130)	(0.134)	(0.181)	(0.103)

Notes: This table reports estimated correlations between peer quality and school treatment effect parameters for Regents math scores. The ATE is a school's average treatment effect, and other treatment effect parameters are school-specific interactions with student characteristics. Estimates come from maximum likelihood models fit to school-specific regression coefficients from a value-added model controlling for observed characteristics.

Table 3.13: Joint Distribution of Peer Quality and Treatment Effect Parameters for PSAT Scores/10

		Peer	Control function parameters								
		quality	ATE	Female	Black	Hispanic	Sub. lunch	Log tract inc.	Math score	Reading score	Pref. coef.
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Mean		0	0	-0.033	-0.284	-0.259	-0.006	-0.005	0.963	1.032	-0.003
		-	-	(0.010)	(0.026)	(0.027)	(0.011)	(0.010)	(0.016)	(0.011)	(0.001)
Standard deviation		0.884	0.401	0.111	0.333	0.352	0.111	0.059	0.240	0.152	0.017
		(0.056)	(0.048)	(0.012)	(0.023)	(0.026)	(0.017)	(0.022)	(0.016)	(0.073)	(0.011)
Correlations:	ATE	0.979									
		(0.086)									
	Female	-0.251	-0.315								
		(0.094)	(0.068)								
	Black	-0.130	-0.253	0.020							
		(0.124)	(0.090)	(0.160)							
	Hispanic	-0.168	-0.274	0.112	0.932						
		(0.094)	(0.079)	(0.150)	(0.123)						
	Subsidized lunch	-0.197	-0.211	0.252	-0.131	-0.120					
		(0.144)	(0.101)	(0.117)	(0.135)	(0.124)					
	Log census tract income	0.198	0.280	-0.228	-0.183	-0.122	-0.545				
		(0.219)	(0.212)	(0.241)	(0.264)	(0.247)	(0.276)				
	Eighth grade math score	0.709	0.701	-0.117	-0.005	-0.090	-0.099	0.022			
		(0.123)	(0.102)	(0.093)	(0.125)	(0.108)	(0.135)	(0.220)			
	Eighth grade reading score	0.164	0.249	-0.219	0.011	-0.084	0.108	0.446	0.246		
		(0.230)	(0.121)	(0.074)	(0.067)	(0.065)	(0.072)	(0.198)	(0.287)		
	Preference coefficient (ψ_j)	0.377	0.291	-0.159	-0.114	-0.062	-0.157	0.334	0.100	-0.109	
		(0.280)	(0.145)	(0.039)	(0.038)	(0.055)	(0.066)	(0.117)	(0.074)	(0.105)	

Notes: This table shows the estimated joint distribution of peer quality and school treatment effect parameters for PSAT scores divded by 10. The ATE is a school's average treatment effect, and other treatment effect parameters are school-specific interactions with student characteristics. Estimates come from maximum likelihood models fit to school-specific regression coefficients from a control function model controlling for observed characteristics, distance to school and unobserved tastes from the choice model.

Table 3.14: Joint Distribution of Peer Quality and Treatment Effect Parameters for High School Graduation

		Peer	Control function parameters								
		quality	ATE	Female	Black	Hispanic	Sub. lunch	Log tract inc.	Math score	Reading score	Pref. coef.
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Mean		0	0	0.063	-0.006	-0.013	-0.013	0.002	0.132	0.062	0.000
		-	-	(0.004)	(0.007)	(0.008)	(0.003)	(0.003)	(0.003)	(0.002)	(0.000)
Standard deviation		0.100	0.043	0.047	0.090	0.103	0.024	0.024	0.034	0.027	0.006
		(0.004)	(0.008)	(0.004)	(0.007)	(0.007)	(0.003)	(0.004)	(0.002)	(0.002)	(0.000)
Correlations:	ATE	0.590									
		(0.106)									
	Female	-0.072	-0.549								
		(0.070)	(0.170)								
	Black	-0.226	-0.296	-0.069							
		(0.069)	(0.195)	(0.142)							
	Hispanic	-0.174	-0.237	-0.078	0.956						
		(0.067)	(0.196)	(0.135)	(0.013)						
	Subsidized lunch	0.169	-0.120	0.119	0.171	0.264					
		(0.096)	(0.238)	(0.169)	(0.180)	(0.176)					
	Log census tract income	0.039	0.032	-0.412	-0.113	-0.168	0.177				
		(0.103)	(0.244)	(0.154)	(0.196)	(0.193)	(0.203)				
	Eighth grade math score	-0.396	-0.619	0.075	-0.168	-0.114	0.051	0.036			
		(0.060)	(0.166)	(0.098)	(0.109)	(0.107)	(0.128)	(0.134)			
	Eighth grade reading score	-0.571	-0.570	-0.125	0.188	0.094	-0.194	0.140	0.475		
		(0.059)	(0.180)	(0.112)	(0.136)	(0.134)	(0.153)	(0.157)	(0.103)		
	Preference coefficient (ψ_j)	0.625	0.437	0.123	-0.110	-0.049	0.021	-0.117	-0.246	-0.470	
		(0.044)	(0.180)	(0.084)	(0.089)	(0.086)	(0.120)	(0.123)	(0.078)	(0.078)	

Notes: This table shows the estimated joint distribution of peer quality and school treatment effect parameters for high school graduation. The ATE is a school's average treatment effect, and other treatment effect parameters are school-specific interactions with student characteristics. Estimates come from maximum likelihood models fit to school-specific regression coefficients from a control function model controlling for observed characteristics, distance to school and unobserved tastes from the choice model.

Table 3.15: Joint Distribution of Peer Quality and Treatment Effect Parameters for College Attendance

		Peer	Control function parameters								
		quality	ATE	Female	Black	Hispanic	Sub. lunch	Log tract inc.	Math score	Reading score	Pref. coef.
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Mean		0	0	0.075	-0.010	-0.011	-0.008	0.002	0.118	0.064	-0.002
		-	-	(0.003)	(0.009)	(0.009)	(0.003)	(0.003)	(0.002)	(0.002)	(0.000)
Standard deviation		0.099	0.053	0.035	0.122	0.120	0.031	0.019	0.030	0.024	0.005
		(0.118)	(0.022)	(0.004)	(0.009)	(0.009)	(0.005)	(0.007)	(0.013)	(0.009)	(0.002)
Correlations:	ATE	0.862									
		(0.158)									
	Female	-0.074	-0.307								
		(0.017)	(0.031)								
	Black	-0.035	-0.455	0.040							
		(0.021)	(0.066)	(0.160)							
	Hispanic	-0.135	-0.471	-0.024	0.947						
		(0.019)	(0.031)	(0.043)	(0.019)						
	Subsidized lunch	0.110	0.235	-0.005	-0.390	-0.339					
		(0.027)	(0.078)	(0.139)	(0.119)	(0.117)					
	Log census tract income	-0.215	0.127	-0.182	-0.722	-0.674	0.316				
		(0.065)	(0.238)	(0.287)	(0.246)	(0.241)	(0.242)				
	Eighth grade math score	-0.204	-0.188	0.265	-0.067	-0.028	0.073	-0.437			
		(0.073)	(0.179)	(0.074)	(0.073)	(0.056)	(0.110)	(0.129)			
	Eighth grade reading score	-0.290	-0.121	-0.131	-0.346	-0.364	-0.198	0.217	0.304		
		(0.112)	(0.197)	(0.078)	(0.083)	(0.082)	(0.105)	(0.219)	(0.171)		
	Preference coefficient (ψ_j)	0.770	0.524	0.144	0.106	0.059	0.003	-0.210	-0.072	-0.314	
		(0.119)	(0.130)	(0.068)	(0.056)	(0.057)	(0.129)	(0.233)	(0.238)	(0.183)	

Notes: This table shows the estimated joint distribution of peer quality and school treatment effect parameters for college attendance. The ATE is a school's average treatment effect, and other treatment effect parameters are school-specific interactions with student characteristics. Estimates come from maximum likelihood models fit to school-specific regression coefficients from a control function model controlling for observed characteristics, distance to school and unobserved tastes from the choice model.

Table 3.16: Joint Distribution of Peer Quality and Treatment Effect Parameters for Log College Quality

		Peer	Control function parameters								
		quality	ATE	Female	Black	Hispanic	Sub. lunch	Log tract inc.	Math score	Reading score	Pref. coef.
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Mean		0	0	0.048	-0.037	-0.035	-0.006	-0.001	0.103	0.058	-0.002
		-	-	(0.002)	(0.006)	(0.006)	(0.002)	(0.002)	(0.002)	(0.002)	(0.000)
Standard deviation		0.097	0.063	0.027	0.081	0.084	0.022	0.013	0.031	0.019	0.004
		(0.078)	(0.017)	(0.003)	(0.006)	(0.006)	(0.004)	(0.004)	(0.010)	(0.006)	(0.004)
Correlations:	ATE	0.931									
		(0.051)									
	Female	0.114	0.084								
		(0.018)	(0.021)								
	Black	-0.065	-0.258	-0.023							
		(0.019)	(0.029)	(0.157)							
	Hispanic	-0.239	-0.354	-0.127	0.946						
		(0.018)	(0.021)	(0.059)	(0.048)						
	Subsidized lunch	-0.063	0.060	0.253	-0.334	-0.208					
		(0.035)	(0.038)	(0.082)	(0.085)	(0.071)					
	Log census tract income	0.030	-0.028	-0.333	-0.529	-0.553	0.036				
		(0.060)	(0.068)	(0.121)	(0.132)	(0.135)	(0.109)				
	Eighth grade math score	0.533	0.728	0.381	-0.143	-0.151	0.146	-0.550			
		(0.078)	(0.063)	(0.054)	(0.072)	(0.040)	(0.066)	(0.151)			
	Eighth grade reading score	0.296	0.479	-0.027	-0.266	-0.275	-0.355	0.089	0.466		
		(0.064)	(0.033)	(0.018)	(0.019)	(0.020)	(0.046)	(0.088)	(0.070)		
	Preference coefficient (ψ_j)	0.750	0.623	0.135	0.033	-0.061	-0.086	0.139	0.310	0.161	
		(0.076)	(0.041)	(0.008)	(0.019)	(0.009)	(0.021)	(0.050)	(0.059)	(0.033)	

Notes: This table shows the estimated joint distribution of peer quality and school treatment effect parameters for college quality. The ATE is a school's average treatment effect, and other treatment effect parameters are school-specific interactions with student characteristics. Estimates come from maximum likelihood models fit to school-specific regression coefficients from a control function model controlling for observed characteristics, distance to school and unobserved tastes from the choice model.

Table 3.17: Preferences, Peer Quality, and Math Effects, Alternative Measures of Popularity

	Log first-choice share		Minus log sum of ranks	
		Control		Control
	Value-added	function	Value-added	function
	(1)	(2)	(3)	(4)
Peer quality	0.487	0.542	0.036	0.038
	(0.071)	(0.062)	(0.005)	(0.005)
ATE	-0.009	-0.034	-0.001	-0.002
	(0.045)	(0.040)	(0.003)	(0.003)
Match effect	-0.091	-0.219	-0.004	-0.012
	(0.043)	(0.047)	(0.003)	(0.004)
N	15892		21684	

Notes: This table reports estimates from regressions of alternative measures of school popularity on peer quality and school effectiveness. The dependent variable in columns (1) and (2) is the log of the share of students in a covariate cell ranking each school first, and the dependent variable in columns (3) and (4) is minus the log of the sum of ranks for students in the cell. Unranked schools are assigned one rank below the least-preferred ranked school. Covariate cells are defined by borough, gender, race, subsidized lunch status, an indicator for students above the median of census tract median income, and tercile of the average of eighth grade math and reading scores. Peer quality is constructed as the average predicted Regents math score for enrolled students. Treatment effect estimates are empirical Bayes posterior mean predictions of Regents math effects. Columns (1) and (3) report results from value-added models, while columns (2) and (4) report results from control function models. All regressions include cell indicators. Standard errors are double-clustered by school and covariate cell.

Table 3.18: Preferences for Peer Quality and Regents Math Effects Among Students Ranking Fewer Than 12 Choices

	Value-added models				Control function models			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Peer quality	0.452		0.487	0.445	0.445		0.491	0.485
	(0.071)		(0.073)	(0.077)	(0.065)		(0.067)	(0.068)
ATE		0.276	-0.052	-0.040		0.250	-0.073	-0.073
		(0.053)	(0.053)	(0.054)		(0.052)	(0.049)	(0.049)
Match effect				-0.092				-0.184
				(0.050)				(0.055)
N	20898							

Notes: This table reports estimates from regressions of school popularity on peer quality and school effectiveness, restricted to the subsample of students who ranked fewer than 12 programs. School popularity is measured as the estimated mean utility for each school and covariate cell in the choice model from Table 4. Covariate cells are defined by borough, gender, race, subsidized lunch status, an indicator for students above the median of census tract median income, and tercile of the average of eighth grade math and reading scores. Peer quality is constructed as the average predicted Regents math score for enrolled students. Treatment effect estimates are empirical Bayes posterior mean predictions of Regents math effects. Mean utilities, peer quality, and treatment effects are scaled in standard deviation units. Columns (1)-(4) report results from value-added models, while columns (5)-(8) report results from control function models. All regressions include cell indicators and weight by the inverse of the squared standard error of the mean utility estimates. Standard errors are double-clustered by school and covariate cell.

Table 3.19: Preferences, Peer Quality, and Math Effects, Alternative Treatment Effect Models

	Matched first choice model				Distance instrument model			
	(1)	(2)	(3)	(4)				
Peer quality	0.367		0.400	0.406	0.397		0.402	0.408
	(0.053)		(0.054)	(0.067)	(0.058)		(0.060)	(0.060)
ATE		0.209	-0.058	-0.036		0.236	-0.009	-0.027
		(0.045)	(0.043)	(0.045)		(0.046)	(0.044)	(0.045)
Match effect				-0.092				-0.129
				(0.049)				(0.041)
N	21684							

Notes: This table reports estimates from regressions of school popularity on peer quality and alternative measures of school effectiveness. Estimates in columns (1)-(4) come from an OLS regression of Regents math scores on school indicators interacted with covariates, with controls for distance and fixed effects for first choice schools. Estimates in columns (5)-(8) come from a regression of Regents math scores on school indicators interacted with covariates and control functions measuring mean preferences for each school, excluding distance controls. School popularity is measured as the estimated mean utility for each school and covariate cell in the choice model from Table 4. Covariate cells are defined by borough, gender, race, subsidized lunch status, an indicator for students above the median of census tract median income, and tercile of the average of eighth grade math and reading scores. Peer quality is constructed as the average predicted Regents math score for enrolled students. Treatment effect estimates are empirical Bayes posterior mean predictions of Regents math effects. Mean utilities, peer quality, and treatment effects are scaled in standard deviation units. All regressions include cell indicators and weight by the inverse of the squared standard error of the mean utility estimates. Standard errors are doubleclustered by school and covariate cell.

Table 3.20: Potential Achievement Gains from Ranking Schools by Effectiveness, by Baseline Test Score Quartile

	Observed rankings			Rankings based on effectiveness			Increase in
	Peer quality	ATE	Match	Peer quality	ATE	Match	effectiveness
Baseline quartile	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Lowest	-0.084	0.015	0.015	0.312	0.452	0.356	0.779
Second	0.011	0.042	0.005	0.395	0.469	0.122	0.545
Third	0.127	0.074	-0.011	0.329	0.464	0.018	0.419
Highest	0.399	0.155	-0.157	0.106	0.324	0.149	0.475

Notes: This table summarizes Regents math score gains that parents could achieve by ranking schools based on effectiveness, separately by baseline math score quartile. Columns (1)-(3) report average peer quality, average treatment effects, and average match effects for schools ranked first by students in each quartile. Columns (4)-(6) display corresponding statistics for hypothetical rankings that list schools in order of their treatment effects. Column (7) reports the difference in treatment effects (ATE+match) between the top-ranked school when rankings are based on effectiveness and the observed top-ranked school. Treatment effect estimates come from control function models. All calculations are restricted to ranked schools within the home borough.

Bibliography

Atila Abdulkadiroğlu, Joshua D. Angrist, Susan M. Dynarski, Thomas J Kane, and Parag A. Pathak. Accountability and flexibility in public schools: Evidence from boston's charters and pilots. *The Quarterly Journal of Economics*, 126(2):699–748, 2011.

Atila Abdulkadiroğlu, Parag A. Pathak, and Christopher R. Walters. Free to choose: can school choice reduce student achievement? *American Economic Journal: Applied Economics*, 10(1):175–206, 2018.

Atila Abdulkadiroğlu, Parag A. Pathak, and Alvin E. Roth. The new york city high school match. *American Economic Review: Papers & Proceedings*, 95(2):364–367, May 2005.

Atila Abdulkadiroğlu, Parag A. Pathak, and Alvin E. Roth. Strategy-proofness versus efficiency in matching with indifferences: redesigning the nyc high school match. *American Economic Review*, 99(5):1954–78, December 2009.

Atila Abdulkadiroğlu, Joshua D. Angrist, and Parag A. Pathak. The elite illusion: achievement effects at boston and new york exam schools. *Econometrica*, 82(1):137–196, 2014.

Atila Abdulkadiroğlu, Nikhil Agarwal, and Parag A. Pathak. The welfare effects of coordinated assignment: evidence from the new york city high school match. *American Economic Review*, 107(12):3635–3689, 2017a.

Atila Abdulkadiroğlu, Parag A. Pathak, Jonathan Schellenberg, and Christopher R. Walters. Do parents value school effectiveness? NBER working paper no. 23912, 2017b.

Nikhil Agarwal and Paulo Somaini. Demand analysis using strategic reports: An application to a school choice mechanism. *Econometrica*, 86(2):391–444, 2018.

Joseph G Altonji, Prashant Bharadwaj, and Fabian Lange. Changes in the characteristics of american youth: Implications for adult outcomes. *Journal of Labor Economics*, 30(4): 783–828, 2012.

Joseph G. Altonji, Ching-I Huang, and Christopher R. Taber. Estimating the cream skimming effect of school choice. *Journal of Political Economy*, 123(2):266–324, 2015.

Michael L Anderson. Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American Statistical Association*, 103(484):1481–1495, 2008.

Joshua D. Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 1 edition, 2009.

Joshua D. Angrist, Guido W Imbens, and Alan B Krueger. Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1):57–67, 1999.

Joshua D. Angrist, Parag A. Pathak., and Christopher R. Walters. Explaining charter school effectiveness. *American Economic Journal: Applied Economics*, 5(4):1–27, 2013.

Joshua D. Angrist, Sarah R. Cohodes, Susan M. Dynarski, Parag A. Pathak, and Christopher R. Walters. Stand and deliver: effects of boston's charter high schools on college preparation, entry and choice. *Journal of Labor Economics*, 34(2):275–318, 2016a.

Joshua D. Angrist, Peter D. Hull, Parag A. Pathak, and Christopher R. Walters. Interpreting tests of school vam validity. *American Economic Review: Papers & Proceedings*, 106(5): 388–392, 2016b.

Joshua D. Angrist, Peter D. Hull, Parag A. Pathak, and Christopher R. Walters. leveraging lotteries for school value-added: Testing and estimation. *The Quarterly Journal of Economics*, 132(4):2061–2062, 2017.

M Caridad Araujo, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady. Teacher quality and learning outcomes in kindergarten. *The Quarterly Journal of Economics*, 131 (3):1415–1453, 2016.

David Arnold, Will Dobbie, and Crystal S Yang. Racial bias in bail decisions. *The Quarterly Journal of Economics*, 133(4):1885–1932, 2018.

Christopher N. Avery and Parag A. Pathak. The distributional consequences of public school choice. NBER Working Paper 21525, 2015.

Andrew Bacher-Hicks, Thomas J Kane, and Douglas O Staiger. Validating teacher effect estimates using changes in teacher assignments in los angeles. Technical report, National Bureau of Economic Research, 2014.

Levon Barseghyan, Damon Clark, and Stephen Coate. Public school choice: an economic analysis. NBER working paper no. 20701, 2014.

Patrick Bayer, Fernando Ferreira, and Robert McMillan. A unified framework for measuring preferences for schools and neighborhoods. *Journal of Political Economy*, 115(4):588–638, 2007.

Stephen B Billings, David J Deming, and Jonah Rockoff. School segregation, educational attainment, and crime: Evidence from the end of busing in charlotte-mecklenburg. *The Quarterly Journal of Economics*, 129(1):435–476, 2013.

Sandra E. Black. Do better schools matter? parental valuation of elementary education. *The Quarterly Journal of Economics*, 14(2):577–599, 1999.

Alan S Blinder. Wage discrimination: reduced form and structural estimates. *Journal of Human Resources*, pages 436–455, 1973.

Richard Blundell and Rosa L. Matzkin. Control functions in nonseparable simultaneous equations models. *Quantitative Economics*, 5(2):271–295, 2014.

Timothy N Bond and Kevin Lang. The black–white education scaled test-score gap in grades k-7. *Journal of Human Resources*, 53(4):891–917, 2018.

Kevin Booker, Tim R. Sass, Brian Gill, and Ron Zimmer. The effects of charter high schools on educational attainment. *Journal of Labor Economics*, 29(2):377–415, 2011.

John Bound and Richard B Freeman. What went wrong? the erosion of relative earnings and employment among young black men in the 1980s. *The Quarterly Journal of Economics*, 107(1):201–232, 1992.

John Bound, David A Jaeger, and Regina M Baker. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical association*, 90(430):443–450, 1995.

Robert Brame, Michael G Turner, Raymond Paternoster, and Shawn D Bushway. Cumulative prevalence of arrest from ages 8 to 23 in a national sample. *Pediatrics*, 129(1):21–27, 2012.

Jennifer Broatch and Sharon Lohr. Multidimensional assessment of value added by teachers to real-world outcomes. *Journal of Educational and Behavioral Statistics*, 37(2):256–277, 2012.

Simon Burgess, Ellen Greaves, Anna Vignoles, and Deborah Wilson. What parents want: school preferences and school choice. *The Economic Journal*, 125(587):1262–1289, 2014.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Robust inference with multiway clustering. *Journal of Business and Economic Statistics*, 29(2):238–249, 2011.

David Card. Using geographic variation in college proximity to estimate the return to schooling. In L.N. Christofides, E.K. Grant, and R. Swidinsky, editors, *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*. University of Toronto Press, Toronto, 1995.

Pedro Carneiro, Claire Crawford, and Alissa Goodman. The impact of early cognitive and non-cognitive skills on later outcomes. 2007.

Elizabeth U Cascio and Douglas O Staiger. Knowledge, tests, and fadeout in educational interventions. Technical report, National Bureau of Economic Research, 2012.

Gary Chamberlain. Analysis of covariance with qualitative data. *The Review of Economic Studies*, 47(1):225–238, 1980.

Gary E Chamberlain. Predictive effects of teachers and schools on test scores, college attendance, and earnings. *Proceedings of the National Academy of Sciences*, 2013. ISSN 0027-8424. doi: 10.1073/pnas.1315746110.

Amitabh Chandra, Amy Finkelstein, Adam Sacarny, and Chad Syverson. Health care exceptionalism? performance and allocation in the us health care sector. *American Economic Review*, 106(8):2110–44, 2016.

Raj Chetty and Nathaniel Hendren. The impacts of neighborhoods on intergenerational mobility ii: county-level estimates. NBER working paper no. 23002, 2017.

Raj Chetty, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. How does your kindergarten classroom affect your earnings? evidence from project star. *The Quarterly Journal of Economics*, 126(4):1593–1660, 2011.

Raj Chetty, John N. Friedman, and Jonah E. Rockoff. Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9): 2593–2632, September 2014a.

Raj Chetty, John N. Friedman, and Jonah E. Rockoff. Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9):2633–79, September 2014b.

Raj Chetty, John N Friedman, and Jonah Rockoff. Using lagged outcomes to evaluate bias in value-added models. *American Economic Review: Papers & Proceedings*, 106(5):393–99, 2016.

Raj Chetty, John N Friedman, and Jonah E Rockoff. Measuring the impacts of teachers: reply. *American Economic Review*, 107(6):1685–1717, 2017a.

Raj Chetty, John N. Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan. Mobility report cards: the role of colleges in intergenerational mobility. The Equality of Opportunity Project, January, 2017b.

John E. Chubb and Terry M. Moe. *Politics, Markets, and America's Schools.* Brookings Institutution Press, Washington, DC, 1990.

Philip J. Cook and Songman Kang. Birthdays, schooling, and crime: Regression-discontinuity analysis of school performance, delinquency, dropout, and crime initiation. *American Economic Journal: Applied Economics*, 8(1):33–57, January 2016.

Thomas Cornelissen, Christian Dustmann, Anna Raute, and Uta Schönberg. Who benefits from universal childcare? estimating marginal returns to early childcare attendance. Working paper, 2016.

Julie B. Cullen, Brian A. Jacob, and Steven D. Levitt. The effect of school choice on participants: evidence from randomized lotteries. *Econometrica*, 74(5):1191–1230, 2006.

Flavio Cunha, James J Heckman, and Susanne M Schennach. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3):883–931, 2010.

Gordon B. Dahl. Mobility and the return to education: testing a roy model with multiple markets. *Econometrica*, 70(6):2367–2420, 2002.

Stacy B. Dale and Alan B. Krueger. Estimating the payoff to attending a more selective college: an application of selection on observables and unobservables. *The Quarterly Journal of Economics*, 117(4):1491–1527, 2002.

Stacy B. Dale and Alan B. Krueger. Estimating the effects of college characteristics over the career using administrative earnings data. *Journal of Human Resources*, 49(2):323–358, 2014.

David Deming. Using school choice lotteries to test measures of school effectiveness. *American Economic Review: Papers & Proceedings*, 104(5):406–411, 2014.

David J. Deming. Better schools, less crime? *The Quarterly Journal of Economics*, 126(4): 2063–2115, 2011.

David J. Deming, Justine S. Hastings, Thomas J. Kane, and Douglas O. Staiger. School choice, school quality, and postsecondary attainment. *American Economic Review*, 104 (3):991–1013, 2014.

Betsy DeVos. Comments at Brookings Institution event on the 2016 Education Choice and Competition Index, March 29th. https://www.brookings.edu/events/the-2016-education-choice-and-competition-index., 2017.

Monica Disare. City to eliminate high school admissions method that favored families with time and resources. Chalkbeat, June 6. Available at: https://www.chalkbeat.org/posts/ny/2017/06/06/city-to-eliminate-high-school-admissions-method-that-favored-families-with-time-and-resources/, Last accessed December 2017, 2017.

Will Dobbie and Roland G. Fryer. The impact of attending a school with high-achieving peers: evidence from the new york city exam schools. *American Economic Journal: Applied Economics*, 6(3):58–75, 2014.

Jeffrey A. Dubin and Daniel L. McFadden. An econometric analysis of residential electric appliance holdings and consumption. *Econometrica*, 52(2):345–362, 1984.

Lester E. Dubins and David A. Freedman. Machiavelli and the gale-shapley algorithm. 88: 485–494, 1981.

Susan M. Dynarski, Joshua Hyman, and Diane Whitmore Schanzenbach. Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion. *Journal of Policy Analysis and Management*, 32(4):692–717, 2013.

Dennis Epple and Richard Romano. Competition between private and public schools, vouchers, and peer-group effects. *American Economic Review*, 62(1):33–62, 1998.

Dennis Epple and Richard Romano. Educational vouchers and cream skimming. *International Economic Review*, 49(4):1395–1435, 2008.

Dennis Epple, David N. Figlio, and Richard Romano. Competition between private and public schools: testing stratification and pricing predictions. *Journal of Public Economics*, 88:1215–1245, 2004.

Gabrielle Fack, Julien Grenet, and Yinghua He. Beyond truth-telling: preference estimation with centralized school choice. Working Paper, Paris School of Economics, 2015.

David N. Figlio and Maurice E. Lucas. What's in a grade? school report cards and the housing market. *American Economic Review*, 94(3):591–604, 2004.

Amy Finkelstein, Matthew Gentzkow, Peter Hull, and Heidi Williams. Adjusting risk adjustment - accounting for variation in diagnostic intensity. *New England Journal of Medicine*, 376(7):608–610, 2017.

Sarah Flèche. Teacher quality, test scores and non-cognitive skills: Evidence from primary school teachers in the uk. 2017.

Nicole Fortin, Thomas Lemieux, and Sergio Firpo. Decomposition methods in economics. In *Handbook of labor economics*, volume 4, pages 1–102. Elsevier, 2011.

Richard B Freeman. The economics of crime. *Handbook of labor economics*, 3:3529–3571, 1999.

Milton Friedman. *Capitalism and Freedom*. Cambridge University Press, 1962.

Roland G Fryer. Racial inequality in the 21st century: The declining significance of discrimination. In *Handbook of Labor Economics*, volume 4, pages 855–971. Elsevier, 2011.

Roland G Fryer. Teacher incentives and student achievement: Evidence from new york city public schools. *Journal of Labor Economics*, 31(2):373–407, 2013.

Roland G. Fryer. An empirical analysis of racial differences in police use of force. *Journal of Political Economy*, forthcoming.

Roland G Fryer and Steven D Levitt. The black-white test score gap through third grade. *American Law and Economics Review*, 8(2):249–281, 2006.

Roland G Fryer and Steven D Levitt. Testing for racial differences in the mental ability of young children. *American Economic Review*, 103(2):981–1005, 2013.

David Gale and Lloyd S. Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

Jonah B Gelbach. When do covariates matter? and which ones, and how much? *Journal of Labor Economics*, 34(2):509–543, 2016.

Seth Gershenson. Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy*, 11(2):125–149, 2016.

Steven Glazerman and Dallas Dotter. Market signals: evidence on the determinants and consequences of school choice from a citywide lottery. Mathematica Policy Research working paper, 2016.

Steven Glazerman, Ali Protik, Bing-ru Teh, Julie Bruch, and Jeffrey Max. Transfer incentives for high-performing teachers: Final results from a multisite randomized experiment. ncee 2014-4004. *National Center for Education Evaluation and Regional Assistance*, 2013.

Cassandra M. Guarino, Mark D. Reckase, and Jeffrey M. Wooldridge. Can value-added measures of teacher performance be trusted? *Education Finance and Policy*, 10(1):117–156, 2015.

Guillaume Haeringer and Flip Klijn. Constrained school choice. *Journal of Economic Theory*, 144(5):1921–1947, 2009.

Eric A. Hanushek. Throwing money at schools. *Journal of Policy Analysis and Management*, 1(1):19–41, 1981.

Eric A Hanushek. The economic value of higher teacher quality. *Economics of Education review*, 30(3):466–479, 2011.

Miles D Harer and Darrell Steffensmeier. The differing effects of economic inequality on black and white rates of violence. *Social Forces*, 70(4):1035–1054, 1992.

Douglas N. Harris and Matthew Larsen. What schools do families want (and why)? Technical report, Education Research Alliance for New Orleans, 2014.

Justine Hastings, Ali Hortačsu, and Chad Syverson. Sales force and competition in financial product markets: The case of mexico's social security privatization. *Econometrica*, 85(6): 1723–1761, 2017.

Justine S. Hastings and Jeffrey M. Weinstein. Information, school choice, and academic achievement: evidence from two experiments. *The Quarterly Journal of Economics*, 123 (4):1373–1414, 2008.

Justine S. Hastings, Thomas J. Kane, and Douglas O. Staiger. Heterogeneous preferences and the efficacy of public school choice. Working paper, 2009.

Jerry A. Hausman and Paul A. Ruud. Specifying and testing econometric models for rank-ordered data. *Journal of Econometrics*, 34(1-2):83–104, 1987.

James Heckman, Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz. Analyzing social experiments as implemented: A reexamination of the evidence from the highscope perry preschool program. *Quantitative Economics*, 1(1):1–46, 2010b.

James J Heckman and Tim Kautz. Hard evidence on soft skills. *Labour economics*, 19(4): 451–464, 2012.

James J. Heckman and Richard Robb. Alternative methods for evaluating the impact of interventions: an overview. *Journal of Applied Econometrics*, 30(1-2):239–267, 1985.

James J Heckman, Jora Stixrud, and Sergio Urzua. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, 24 (3):411–482, 2006a.

James J. Heckman, Sergio Urzua, and Edward Vytlacil. Understanding instrumental variables estimates in models with essential heterogeneity. *Review of Economics and Statistics*, 88(3):389–432, 2006b.

James J. Heckman, Sergio Urzua, and Edward Vytlacil. Instrumental variables in models with multiple outcomes: the general unordered case. *Annales d'Economie et de Statistique*, 91-92:151–174, 2008.

James J Heckman, Seong Hyeok Moon, Rodrigo Pinto, Peter A Savelyev, and Adam Yavitz. The rate of return to the highscope perry preschool program. *Journal of Public Economics*, 94(1):114–128, 2010a.

Stephen B Holt and Seth Gershenson. The impact of demographic representation on absences and suspensions. *Policy Studies Journal*, 2017.

Caroline M. Hoxby. Does competition among public schools benefit students and taxpayers? *American Economic Review*, 90(5):1209–1238, 2000.

Caroline M. Hoxby. School choice and school productivity: could school choice be a tide that lifts all boats? In Caroline M. Hoxby, editor, *The Economics of School Choice*. University of Chicago Press, Chicago, IL, 2003.

Chang Hsieh and Miguel Urquiola. The effects of generalized school choice on achievement and stratification: evidence from chile's voucher program. *Journal of Public Economics*, 90:1477–1503, 2006.

Peter D. Hull. Estimating hospital quality with quasi-experimental data. Working paper, 2016.

Scott A. Imberman and Michael F. Lovenheim. Does the market value value-added? evidence from housing prices after a public release of school and teacher value-added. *Journal of Urban Economics*, 91:104–121, 2016.

C Kirabo Jackson. Teacher quality at the high school level: The importance of accounting for tracks. *Journal of Labor Economics*, 32(4):645–684, 2014.

C Kirabo Jackson. What do test scores miss? the importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, 126(5):2072–2107, 2018.

C Kirabo Jackson, Jonah E Rockoff, and Douglas O Staiger. Teacher effects and teacher-related policies. *Annual Review of Economics*, 6(1):801–825, 2014.

Brian A Jacob and Lars Lefgren. Are idle hands the devil's workshop? incapacitation, concentration, and juvenile crime. *American Economic Review*, 93(5):1560–1577, 2003.

Brian A. Jacob and Lars Lefgren. What do parents value in education? an empirical investigation of parents' revealed preferences for teachers. *The Quarterly Journal of Economics*, 122(4):1603–1637, 2007.

Brian A. Jacob and Lars Lefgren. Principals as agents: subjective performance assessment in education. *Journal of Labor Economics*, 26(1):101–136, 2008.

Brian A. Jacob, Lars Lefgren, and David P. Sims. The persistence of teacher-induced learning. *Journal of Human Resources*, 45(4):915–943, 2010.

Christopher Jencks and Meredith Phillips. The black-white test score gap. 1998.

Frank Lancaster Jones. On decomposing the wage gap: a critical comment on blinder's method. *The Journal of Human Resources*, 18(1):126–130, 1983.

Thomas J. Kane and Douglas O. Staiger. The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives*, 16(4):91–114, 2002.

Thomas J Kane and Douglas O Staiger. Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research, 2008. NBER working paper no. 14607.

Thomas J. Kane, Daniel F. McCaffrey, and Douglas O. Staiger. Have we identified effective teachers? validating measures of effective teaching using random assignment. *Gates Foundation Report*, 2013.

Adam Kapor, Christopher A. Neilson, and Seth D. Zimmerman. Heterogeneous beliefs and school choice mechanisms. Working paper, 2017.

Lars J. Kirkebøen, Edwin Leuven, and Magne Mogstad. Field of study, earnings, and self-selection. *The Quarterly Journal of Economics*, 131(3):1057–1111, 2016.

Patrick Kline. Oaxaca-blinder as a reweighting estimator. *American Economic Review: Papers & Proceedings*, 101(3):532–537, 2011.

Patrick Kline and Andres Santos. A score based approach to wild bootstrap inference. *Journal of Econometric Methods*, 1(1):23–41, 2012.

Patrick Kline and Christopher R. Walters. Evaluating public programs with close substitutes: the case of head start. *The Quarterly Journal of Economics*, 131(4):1795–1848, 2016.

Cory Koedel, Kata Mihaly, and Jonah E. Rockoff. Value-added modeling: a review. *Economics of Education Review*, 47:180–195, 2015.

Helen F. Ladd. *Market-based Reforms in Urban Education.* Economic Policy Institute, Washington, DC, 2002.

Kevin Lang and Michael Manove. Education and labor market discrimination. *American Economic Review*, 101(4):1467–96, June 2011.

Ashley Langer. (dis)incentives for demographic price discrimination in the new vehicle market. Working paper, 2016.

Lung-Fei Lee. Generalized econometric models with selectivity. *Econometrica*, 51(2):507–512, 1983.

Mary Kay Linge and Joshua Tanzer. The top 40 public high schools in nyc. New York Post, September 17, 2016.

Allen E Liska and Paul E Bellair. Violent-crime rates and racial composition: Covergence over time. *American journal of sociology*, 101(3):578–610, 1995.

Lance Lochner. Non-production benefits of education: Crime, health and good citizenship. In Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, editors, *Handbook of the Economics of Education*, volume 5. Elsevier, Oxford, 2011.

Lance Lochner and Enrico Moretti. The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American Economic Review*, 94(1):155–189, March 2004. doi: 10.1257/000282804322970751. URL `http://www.aeaweb.org/articles?id=10.1257/000282804322970751`.

W. Bentley MacLeod and Miguel Urquiola. Reputation and school competition. *American Economic Review*, 105(11):3471–3488, 2015.

W. Bentley MacLeod, Evan Riehl, Juan E. Saavedra, and Miguel Urquiola. The big sort: college reputation and labor market outcomes. *American Economic Journal: Applied Economics*, 9(3):223–261, 2017.

Charles F Manski. Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542, 1993.

Terrie E. Moffitt, Louise Arseneault, Daniel Belsky, Nigel Dickson, Robert J. Hancox, HonaLee Harrington, Renate Houts, Richie Poulton, Brent W. Roberts, Stephen Ross, Malcolm R. Sears, W. Murray Thomson, and Avshalom Caspi. A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences*, 108(7):2693–2698, 2011.

Carl N. Morris. Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55, 1983.

Jack Mountjoy. Community colleges and upward mobility. Working paper, 2017.

Lori Nathanson, Sean Corcoran, and Christine Baker-Smith. High school choice in new york city: a report on the school choices and placements of low-achieveing students. Research Alliance for New York City Schools, April, 2013.

Derek Neal. The effects of catholic secondary schooling on educational achievement. *Journal of Labor Economics*, 15(1):98–123, 1997.

Derek Neal. Why has black–white skill convergence stopped? *Handbook of the Economics of Education*, 1:511–576, 2006.

Derek Neal. The design of performance pay in education. In *Handbook of the Economics of Education*, volume 4, pages 495–550. Elsevier, 2011.

Derek A Neal and William R Johnson. The role of premarket factors in black-white wage differences. *Journal of Political Economy*, 104(5):869–895, 1996.

New York City Department of Education. Directory of the new york city public high schools, 2003-2004. New York, New York, 2003.

New York City Department of Education. 2003-2004 annual school reports. New York, New York, 2004.

New York City Department of Education. 2016-17 school quality reports. http://schools.nyc.gov/Accountability/tools/report/default.htm. Accessed December 11, 2017., 2017.

Ronald Oaxaca. Male-female wage differentials in urban labor markets. *International economic review*, pages 693–709, 1973.

Ronald L Oaxaca and Michael Ransom. Calculation of approximate variances for wage decomposition differentials. *Journal of Economic and Social Measurement*, 24(1):55–61, 1998.

Ronald L Oaxaca and Michael R Ransom. Identification in detailed wage decompositions. *Review of Economics and Statistics*, 81(1):154–157, 1999.

June O'Neill. The role of human capital in earnings differences between black and white men. *Journal of economic Perspectives*, 4(4):25–45, 1990.

Parag A. Pathak and Tayfun Sönmez. School admissions reform in chicago and england: comparing mechanisms by their vulnerability to manipulation. *American Economic Review*, 103(1):80–106, 2013.

Nathan Petek and Nolan Pope. The multidimensional impact of teachers on students. Technical report, University of Chicago Working Paper, 2016.

Michael J Podgursky and Matthew G Springer. Teacher performance pay: A review. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, 26(4):909–950, 2007.

Steven Raphael. The socioeconomic status of black males: The increasing importance of incarceration. *Public policy and the income distribution*, pages 319–358, 2006.

Steven Raphael and Sandra V Rozo. Racial disparities in the acquisition of juvenile arrest records. *Journal of Labor Economics*, 37(S1):S125–S159, 2019.

Arthur J. Reynolds, Judy A. Temple, and Suh-Ruu Ou. Preschool education, educational attainment, and crime prevention: Contributions of cognitive and non-cognitive skills. *Children and Youth Services Review*, 32(8):1054 – 1063, 2010. ISSN 0190-7409. Early Childhood to Young Adulthood: Intervention and Alterable Influences on Well-Being.

Steven G Rivkin, Eric A Hanushek, and John F Kain. Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458, 2005.

Herbert Robbins. An empirical bayes approach to statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1:157–163, 1956.

Jonah E Rockoff. The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 94(2):247–252, 2004.

Evan K. Rose and Yotam Shem-Tov. Does incarceration increase crime? Working Paper, 2018.

Alvin E. Roth. The economics of matching: stability and incentives. *Mathematics of Operations Research*, 7:617–628, 1982.

Jesse Rothstein. Good principals or good peers? parental valuation of school characteristics, tiebout equilibrium, and the incentive effects of competition among jurisdictions. *American Economic Review*, 96(4):1333–1350, 2006.

Jesse Rothstein. Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, 125(1):175–214, 2010.

Jesse Rothstein. Measuring the impacts of teachers: comment. *American Economic Review*, 107(6):1656–1684, 2017.

A.D. Roy. Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3(2): 135–146, 1951.

Edward S Shihadeh and Nicole Flynn. Segregation and crime: the effect of black social isolation on the rates of black urban violence. *Social Forces*, 74(4):1325–1352, 1996.

Petra E. Todd and Kenneth I. Wolpin. On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113(485):F3–F33, 2003.

Christopher R Walters. The demand for effective charter schools. *Journal of Political Economy*, 126(6):2179–2223, 2018.

Bruce Western and Becky Pettit. Beyond crime and punishment: Prisons and inequality. *Contexts*, 1(3):37–43, 2002.

Jeffrey M. Wooldridge. Control function methods in applied econometrics. *Journal of Human Resources*, 50(2):420–445, 2015.