ESSAYS IN APPLIED MICROECONOMICS

A Dissertation

by

CHANDON RAE'MEL ADGER

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Steve Puller
Committee Members,	Mark Hoekstra
	Fernando Luco Echeverria
	Joanna Lahey
Head of Department,	Steve Puller

May 2023

Major Subject: Economics

ABSTRACT

In this dissertation, I use a variety of econometric techniques to answer pressing questions about various questions in Applied Microeconomics that are relevant to public policy.

In the first chapter, we study how schools respond to changes in voucher policy. Using quasi-experimental and empirical Industrial Organization techniques, We find that public schools increase school quality while private schools reduce school quality following the expansion of school choice programs. Moreover, the incentive for schools to provide quality varies non-linearly with voucher expansion and depends crucially on both the voucher value as well as the types of students eligible to receive the voucher.

In the second chapter, we study the impact that field training has on a new officer's subsequent policing behavior. Leveraging a setting with as good as random assignment of Field Training officers to new recruits, we find that a one standard deviation increase in a field training officer's propensity to use force (138 percent) is associated with a 12 percent increase in their recruit's subsequent propensity to use force. The effect of having a more aggressive field training officer persists for as much as two and a half years after the recruit completes training.

DEDICATION

To my family and friends, specifically, my mom, who helped me stay level-headed throughout my studies.

ACKNOWLEDGMENTS

I am grateful to my wonderful co-authors Brianna Felegi (Chapter I) and Matthew Ross & CarlyWill Sloan (Chapter II).

I would like to thank my advisor and all of my committee members for sitting on countless meetings to get me to this point.

A special thanks to my countless mentors and peers who have given me advice both scholastically and personally throughout my studies.

Chapter I of this dissertation was supported by Notre Dame's Center for Research on Educational Opportunity (CREO) and the Institute of Educational Initiatives. I am grateful to the Indiana Department of Education for providing access to state administrative records and for supporting independent analyses. I am also grateful to Roberto Peñaloza, who helped organize the data for Chapter I.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor(s) Steve Puller, Mark Hoekstra and Fernando Luco of the Department of Economics and Professor Joanna Lahey of the Bush School of Government & Public Service.

The data analyzed for Chapter I was provided by Notre Dame's Center for Research on Educational Opportunity (CREO) and the Institute of Educational Initiatives in conjunction with the Indiana Department of Education.

The analyses depicted in Chapter I and II were conducted together with Brianna Felegi and Matthew Ross & CarlyWill Sloan, respectively. All other work conducted for the dissertation was completed by myself.

Funding Sources

My graduate studies was supported by a fellowship from Texas A&M University. Additional funding was provided by the Private Enterprise Research Center.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. SUPPLY SIDE RESPONSES IN SCHOOL CHOICE

Despite the growing size of private-school voucher programs, our understanding of their effectiveness relies on results from small-scale randomized control trials. In this paper, we show that those results may not translate to programs at scale by examining changes in school quality following the implementation of the largest voucher program in the United States, the Indiana Choice Scholarship Program. We find that public schools facing high exposure to the policy increased their quality while participating private schools decreased their quality. Public schools with below-median baseline school value-added drive our results, suggesting that the gap in public school quality is shrinking because of the program. We explore these effects in a model of household demand for schools. We show that the incentive to provide quality is nonlinear. Voucher programs only threaten public school enrollment when the voucher amount is large or when a significant proportion of students are eligible to participate. Policymakers interested in adopting and expanding these programs should consider these indirect and nonlinear effects to understand vouchers' impact on educational outcomes.

This paper was supported by Notre Dame's Center for Research on Educational Opportunity (CREO) and the Institute of Educational Initiatives. We are grateful to the Indiana Department of Education for providing access to state administrative records and for supporting independent analyses. We are also grateful to Roberto Peñaloza, who helped organize the data. We thank those at the University of Notre Dame, Texas A&M University, Virginia Tech and the Federal Reserve Bank of Chicago for their insightful feedback. All opinions expressed in this paper represent those of the authors and not necessarily the institutions with which they are affiliated. All errors in this paper are solely the responsibility of the authors.

I Introduction

School choice programs have become a popular tool to eliminate inequities in access to schooling. Private-school vouchers have drawn increasing attention in this effort. In the last 20 years, the number of state-funded voucher programs has increased five-fold, from 5 in 2000 to 27 in 2021. Furthermore, the scale of these programs has grown significantly over time. The first U.S. voucher program, Milwaukee Parental Choice, featured an enrollment limit of 1% of the public school population when it launched in 1991. Today, the average voucher program has no enrollment cap, and around 26% of families qualify to participate [EdChoice, 2021). Moreover, in states that have these programs, nearly 1 in 10 private school students now use a voucher to attend [EdChoice, 2021, National Center for Education Statistics, 2019).

Despite increases in the size of voucher programs, the literature evaluating their effectiveness has relied on small-scale randomized control trials (RCTs) comparing the outcomes of those offered a voucher to those in the control group for a small subset of the total student population [Mayer et al., 2002, Howell et al., 2002, Wolf et al., 2010, Witte et al., 2014, Abdulkadiroğlu et al., 2018).[1] While these RCTs provide useful estimates of the average effect of being offered a voucher, their results may not capture the overall impact of voucher programs when vouchers are implemented on a larger scale. Specifically, economic theory predicts that as these programs expand, schools have the incentive to respond [Friedman, 1962, Chakrabarti, 2008). Examining school responses to voucher programs is essential to understanding how such programs impact educational outcomes.

In this paper, we quantify schools' responses by examining changes in school quality following the adoption of the largest voucher program in the United States. Our context centers around the Indiana Choice Scholarship Program (ICSP), which was initially adopted in 2011 and expanded in 2013. We begin with student-level testing data that covers all students in the state between the 2005-2006 and 2017-2018 academic years (AY). We use these data to construct school-level measures of quality by estimating value-added for both public and private schools. We then identify

[1]See Epple et al. [2017) and Rouse and Barrow [2009) for excellent reviews on the topic.

schools facing greater exposure to the voucher policy by calculating the radial distances between each public and private school within the state. Specifically, we distinguish high-exposure public schools as those that face increased competition because they are located within five miles of a private school. Private schools that accept voucher students are said to face high exposure to the policy if they are in the top tercile of the distribution of the number of public schools within five miles.[2] The resulting data sets track school quality for those in our high-exposure and control groups, both before and after the implementation of the voucher program.[3]

Using these datasets, we estimate the causal effects of the implementation of ICSP on schools using a standard difference-in-differences model. Specifically, we compare the change in school value-added in the years before and after the implementation of the policy for schools facing high exposure to the policy versus those in the control group. Our primary analysis focuses on public schools. We find that, on average, public schools facing the threat of voucher competition saw a statistically significant increase of 0.023 of a standard deviation (s.d.) in their overall school value-added, an increase of 0.03 s.d. in their math value-added, and an increase of 0.013 in their reading value-added. However, improvements in value-added varied within the high-exposure group. Public schools facing the threat of competition and an above-median share of students qualifying for a 90% voucher witnessed the largest improvements in school quality. Specifically, these schools saw increases of 0.039 s.d. in overall school VA, 0.047 s.d. in math VA, and 0.028 s.d. in reading VA.

We further explore the impacts of ICSP on public-school quality by disaggregating the results by several baseline characteristics. Specifically, we examine whether the changes in public school quality differ across schools above/below the median in baseline enrollment, overall school value-added, and income of the census block group where the school is located. While we find no evidence of heterogeneous results across enrollment or household income, high exposure public schools with an above-median baseline school value-added saw almost no changes in our outcomes

[2]The definition of "high exposure" shifts between public and private schools because while only half of the public schools have a private school within five miles, 98% of private schools have a public school within five miles. In Section VI, we discuss an alternative method for distinguishing high exposure private schools.

[3]We separate the analyses of public and private schools.

of interest. This result suggests that initially poor-performing public schools facing the potential threat of competition drive the changes we see in quality. Together, our results lead us to conclude that the gap in public-school quality shrunk following the implementation of ICSP.

We also employ an event-study specification that allows us to examine whether the adoption and expansion of ICSP had differential impacts on public school quality. We find that the adoption of the policy did not elicit differential changes across high exposure and control public schools in our outcome measures of interest. Instead, increases in school quality among high-exposure public schools are seen only after the program's expansion. This result indicates that despite facing potential enrollment losses when the program was adopted, public schools only responded once there was a threat that a majority of their students could leave. We take these results as evidence that the total effect of voucher programs at scale may be very different from the partial equilibrium results found in the existing literature.

To understand how high exposure public schools increase quality, as measured by VA, we combine a school-level dataset on available teachers between the 2010-2011 and 2017-2018 academic years with the National Center for Education Statistics's Common Core of Data on Indiana public schools. We do not find strong evidence that following the implementation of ICSP, high-exposure public schools saw changes in their student-teacher ratios. However, high-exposure public schools saw an increase of 0.7 teachers with a graduate degree and 1.5 teachers with a high-quality certification when compared to the set of control schools. We also find that after the adoption of ICSP, high-exposure public schools saw increases in their attendance and decreases in the percent of students ever suspended or expelled. These findings suggest that in response to ICSP, schools increased quality and made improvements on both the cognitive and non-cognitive dimensions.

Given our results, we pay particular attention to the possibility that changes in the composition of students could generate our findings. To address this concern, we first document the extent to which student sorting occurs after the implementation of ICSP. We find that high-exposure public schools see a decline of 2.8 percentage points (p.p.) in the number of White students and a rise of 2.3 p.p. in the number of Hispanic students after the policy is adopted. We also find that students who use a voucher have slightly higher achievement levels than those who qualify for the voucher, but remain in the public school system.[4] To understand whether these demographic changes drive our results, we run a difference-in-differences specification using predicted value-added. We find that based only on changes in observable characteristics, high-exposure public schools were predicted to see declines in their school value-added. We take these results as evidence that the improvements in school quality are not due to student sorting. We address concerns over non-random student sorting on unobservable characteristics by highlighting the advantages of our value-added estimates since they control for prior achievement. Assuming that prior achievement fully proxies for inputs that affect a student's achievement prior to using the voucher and those inputs are correlated with a student's likelihood of using a voucher, we can mitigate the concerns of this type of sorting.

Our public school results are robust to model specification choices and the adoption of other policy interventions that could threaten the validity of our findings. First, in our event-study specifications, high-exposure public schools and those in the control group appear to have similar trends in school value-added in all years prior to the program, suggesting that our results are not driven by differential trends between the two groups of public schools. We also show that our results are robust to placebo adoption years. Specifically, we re-run our event-study specification, this time setting the adoption of ICSP to two years before the program started. We find that changes in school quality occurred only after the expansion of the voucher policy, further bolstering our conclusions that pre-trends do not drive our results. To address the concern that high-exposure public schools may be concentrated in a small number of urban districts, we run our results dropping each county in Indiana. The analysis produces similar results to the entire state sample. Lastly, we argue that no meaningful policy changes were adopted that would have differentially impacted our two sets of schools and influenced our findings.

For private schools accepting voucher students (from hereon called choice schools), we use

[4]This phenomenon is often referred to as "cream-skimming" and is one of the main critiques of private school voucher programs. However, our results suggest that high-exposure public schools improve their quality despite this sorting on ability.

the constructed dataset to present evidence on their responses to the policy. Choice schools see declines in average quality on all dimensions during the first year of the voucher program. In our difference-in-differences specification, we compare choice schools surrounded by many public schools to those with fewer options to attract students. We find evidence that high-exposure choice schools saw larger decreases in school quality compared to the control group. We discuss the possible theoretical reasons behind the decrease in quality in the following sections. We use the Private School Universe Survey to understand to what extent choice schools alter their school inputs durng our sample period. Following the adoption of ICSP, high-exposure choice schools see a statistically significant increase of 0.93 (off a base mean of 14.24) in their student-teacher ratios. We also find suggestive evidence that control choice schools increase their instructional time to catch up to high-exposure choice schools once the program is adopted.

The assumption underpinning our reduced-form analysis is that potential changes in enrollment create an incentive for schools to change their quality. However, the results cannot quantify this threat or whether it varies depending on program design. We therefore develop a simple demand model allowing us to quantify public schools' exact loss and private schools' exact gain in market share due to ICSP absent any change in quality. We frame our model around the intuition that holding all else constant, increasing voucher eligibility or the voucher amount will reduce product differentiation between the two sets of schools in the market. The reduction in product differentiation reduces the market power of public schools – for which there are now better substitutes – and provides an incentive for them to improve quality to avoid the total loss of market share and corresponding revenue. For the opposite reason, private schools are less incentivized to provide quality. The advantage of our model is that we can estimate the relationship between threats to enrollment and the incentive to change quality for the full range of potential voucher programs with respect to eligibility and voucher size.

Our model provides three important insights: First, small expansions of a voucher program, in terms of either changes in the voucher amount or eligibility, do not threaten public schools' enrollment or influence incentives to provide quality. Second, higher income households value quality significantly more than lower income households, suggesting that voucher programs targeting the latter group will affect quality differently than those including both groups. Third, voucher policies can change competition on two dimensions, depending on their design. When these programs target individuals more likely to attend a public school (absent the policy), competition is focused between public and private schools. When they are expanded to include students likely to always attend a private school, competition is focused between private schools. This shift in competition creates nonlinearities in schools' incentives to provide quality. Specifically, our results show that public (private) schools only have the incentive to increase (decrease) quality in response to potential changes in enrollment only up to a point. Further expansions reduce (increase) the competition faced by public (private) schools and thus reduce (increase) their incentive to provide quality.

Our paper contributes to the growing economics literature on school choice programs. Many papers specifically examining private-school vouchers focus on the direct impact of these policies on the educational outcomes of students offered to participate. One set of papers examines whether participating students experience test score gains [Rouse, 1998, Mayer et al., 2002, Howell et al., 2002, Witte et al., 2014, Wolf et al., 2010, Abdulkadiroğlu et al., 2018, Waddington and Berends, 2018), and [Chingos and Peterson, 2015) focuses on the longer-term educational impacts including high school graduation and college enrollment. Our paper complements this prior work by demonstrating that voucher policies implemented at scale affect the educational outcomes of students not participating in the program. Specifically, we show that as ICSP is implemented, both students remaining in the public school system and those continuing in private schools experience changes in school quality. By establishing these indirect effects of ICSP, we can better understand the total effect of voucher policies as they are adopted and expanded.[5]

A large body of work evaluates the supply-side responses to school choice programs. A majority of these papers focus on the public school response to the introduction of charter schools [Cohodes and Parham, 2021, Imberman, 2011a, Gilraine et al., 2021).[6] We give two reasons why

[5]More broadly, this paper contributes to the literature that calls attention to the limitations of randomized control trials [Lise et al., 2004, Heckman, 1991, Deaton and Cartwright, 2018, Al-Ubaydli et al., 2017).

[6]See Epple et al. [2016) for an excellent review of the effects of charters schools on public school performance.

understanding voucher programs' specific effects are important. First, current policy discussions often center around the adoption and expansion of voucher policies in particular.[7] Second, our results show that ICSP induces changes in quality for both public and participating private schools, suggesting that the effect of voucher policies may differ from the introduction of charter schools.

The most similar work to ours examines the response of public schools to the voucher programs in Milwaukee, Ohio, and Florida [Hoxby, 2003, Figlio and Rouse, 2006, Chakrabarti, 2008, 2013, Rouse et al., 2013, Chiang, 2009, Greene and Marsh, 2009, Figlio and Hart, 2014).[8] Overall, these studies investigate the introduction of school voucher programs and find modest positive effects on public school performance. Our context has attractive empirical properties that allow us to avoid some of the identification issues present within the literature. For example, several papers rely on changes in the degree of private school supply for identification, which may be endogenous to public performance. Other papers identify the effects of voucher programs by leveraging policies that automatically allow students to qualify if their school receives a repeat "F" grade, and the researchers cannot disentangle the effects of school vouchers from the performance effects of accountability pressure. Furthermore, we develop a model that both tests the commonly hypothesized mechanism for why schools respond and allows us to extrapolate our reduced-form results to various voucher designs. To our knowledge, the strategy has not been used in the previous literature because of the extensive data and student tracking requirements.

Figlio et al. [2020) also studies the effects of voucher program expansion by leveraging the Florida Tax Credit Scholarship's growth from 2003 to 2018. The authors use variation in the growth of the program and pre-policy levels of local competition to estimate the intensive margin effects of increased competition on public school performance. They find that students in public schools that faced a higher initial level of competitive pressure saw greater gains in test scores

[7] Since 2021, policymakers from Oklahoma, Nevada, Texas, and Florida have made public announcements supporting the introduction or expansion of voucher policies.

[8] There are several studies examining the specific effect of voucher policies on schools in countries outside of the United States [Hsieh and Urquiola, 2006, Neilson, 2021, Böhlmark and Lindahl, 2015, Muralidharan and Sundararaman, 2015) These papers are similar to ours in that they study programs that serve larger shares of the total student population. However, we might expect different school responses in our context based on differences in baseline private school enrollment and voucher design.

as the program matured. We build on their results in several ways, beginning with our identification strategy. Rather than rely on incremental changes in realized voucher enrollment,[9] our results are estimated off legislated changes in eligibility. Understanding the effects based on this dimension may be of particular interest to policymakers as they can set the limits for eligibility and voucher amount and cannot directly control the number of students participating.[10] Furthermore, our model's estimates show that changes in voucher eligibility and amount can have important implications for whether and to what extent schools respond. We also examine changes in the quality of participating private schools, which is critical for examining voucher programs' total impact. To the best of our knowledge, we are the first to examine changes in private school quality in response to a voucher program within the United States.[11]

The remainder of this paper is organized as follows. In Section II, we provide background information on the Indiana Choice Scholarship Program. Section III summarizes the data used in this paper and describes our constructed measures of school quality and exposure to the policy. Section IV describes the reduced-form empirical strategy and lays out the regression specifications. Section V contains the main results on public schools, which include our heterogeneity analysis, discussion on student sorting, our validity checks, and a discussion on possible mechanisms. Section VI contains the results for choice schools. Section VII discusses our model, and Section VII offers conclusions from this research.

II The Indiana Choice Scholarship Program

The Indiana Choice Scholarship Program (ICSP) is the most expansive single voucher program in the United States in terms of both participation (36,290 participants) and eligibility (over 79% of

[9]Figlio et al. [2020) uses several measures of growth in their analysis. Their preferred specification relies on the log number of scholarship enrollments.

[10]Our results have a slightly different interpretation than those in Figlio et al. [2020) Their results combine the effects of FTC scaling up and maturing over time. Our analysis centers around the first five years after ICSP was expanded, so maturation effects may be less apparent in our context.

[11]Private school responses to voucher programs in the United States is an understudied area. Some papers have studied the effects of these policies on private school enrollment, finances, and school inputs; however, the question of whether and to what extent schools alter their quality is still an open question [Hungerman and Rinz, 2016, Hungerman et al., 2019, Rinz, 2015).

families with children are eligible)[12]. Initially, the program capped participation at 5,000 and 7,500 students for the 2011-2012 and 2012-2013 AYs, respectively. The expansion of ICSP at the start of the 2013-2014 AY eliminated participation caps. Since the expansion, a student can participate in ICSP if they meet the income requirements and qualify under one of eight eligibility tracks.[13]

Income eligibility for vouchers is based on household size and is set as a percentage of the amount to qualify for the Federal Free or Reduced-Price Lunch (FRPL) Program. Students at or below the threshold for FRPL are eligible for a voucher of value up to 90% of per-pupil state funding, while students at or below 300% of the threshold for FRPL are eligible for a voucher of value up to 50% of per-pupil state funding [Indiana Department of Education, 2021b). The actual voucher amount equals the minimum of school tuition and fees or the qualified voucher amount. During the 2020-2021 school year, the average voucher amount for students in grades 1-8 was $5,311 for students qualifying for the 90% voucher and $3,094 for those receiving the 50% voucher ($\leq$ 50% of per-pupil public spending) [Indiana Department of Education, 2021a).

For a student to receive a voucher, they must apply and be accepted into a participating choice school. The choice scholarship application is then completed by a parent (or legal guardian) and submitted by the private school. If a student is awarded a voucher, that money goes directly to the school, and only an award letter detailing the approved amount of the voucher is given to parents.[14] ICSP vouchers are meant to cover tuition and fees at eligible private schools; however, these schools are allowed to charge additional tuition above the voucher amount so long as they are the same charges non-choice-eligible students pay.

The inclusion of both low- and modest-income families makes ICSP unique. The income eligibility threshold for the 2022-2023 academic year in Indiana is about 1.5 times that of the Florida voucher program (Fla. Stat. § 1002.394); 1.85 times higher than that of the programs in

[12]There are three main types of voucher programs including tax credit scholarships, education savings accounts and standard private school voucher programs. The Indiana Choice Scholarship Program is the largest standard private school voucher program. Indiana currently ranks sixth in terms of percentage of current educational expenditures spent on voucher programs.

[13]Information on available tracks can be found on the IDOE website [Indiana Department of Education, 2021c).

[14]The distribution of funding to schools rather than households distinguishes ICSP from tax-credit voucher programs or educational savings accounts, which have also become popular over the last 20 years.

Milwaukee (Wis. Stat. §§ 119.23 and 235), Racine, (Wis. Stat. § 118.60), and Washington, D.C. (DC ST § 38-1853); and about 2.2 times higher than the program in New Orleans (La. Rev. Stat. §§ 17:4011 through 4025). This higher income threshold places additional pressure on the public schools of Indiana. Over 79% of public-school students qualify for a voucher, and participation is not capped at a percentage of public-school enrollment as seen in other voucher programs, suggesting that Indiana is a context where we might expect to see larger impacts on school quality.

III Data

The data for this project come from the Indiana Department of Education (IDOE) through a data agreement with the Center of Research on Educational Opportunity (CREO) at the University of Notre Dame. The IDOE-CREO database contains student-level data with information on the membership, test scores, voucher take-up, and demographics of all students enrolled in a public, private, or charter school in Indiana.[15] The database covers the 2005-2006 through 2017-2018 AY. We focus on students in schools that serve anyone in grades 3-8. Standardized testing is consistent between these grades and is required in both public and private schools in order to remain accredited (Indiana Code §20-32-5-17), which allows for a consistent sample across the sample years. Our dataset is advantageous because it includes information on private schools before ICSP was adopted. Many voucher programs require participating private schools to administer state exams once they accept voucher students, but this means testing data only exists in post-adoption. Indiana private schools had the incentive to be accredited before ICSP because it was required if a school wanted to participate in the Indiana Athletic Association [Indiana Athletic Association, 2021). [16]

Demographic information in the IDOE-CREO database varies depending on whether a student attends a public, private non-choice, or private choice school. (hereafter referred to as public, private, and choice schools, respectively). For all students, we have information on race, age, date of birth, free or reduced-price lunch status, Section 504 status, zoned school district, and standardized testing accommodations. For students attending either public or private schools, we

[15] We focus on public and private school students in this paper.

[16] Another advantage of our dataset is that we can observe students not born in the state of Indiana. Figlio et al. [2020) is restricted to conduct the analysis on students born in the state of Florida.

only have information on whether a student would qualify for a 90% voucher as it is the same cutoff for free/reduced-price lunch. We do not observe whether a student surpasses the cutoff for a 50% voucher. We have additional information on students that use a voucher to attend a private school. Specifically, we also have information on these students' home addresses, the tuition they are charged, their voucher status (50% or 90%), and the amount of the voucher they receive.

We also have access to school directories that outline basic information about the schools in Indiana. This includes data on the opening and closing (if applicable) dates, addresses, school type (public, private, or charter), and lowest/highest grades offered. We construct school-level test scores and demographic information by aggregating individual-level data from students attending each school. Schools must have non-missing test score data for each of the academic years between 2005 and 2017 to be included in the sample. After this restriction, 1,279 public elementary and middle schools and 178 choice schools remain.[17]

We create two other school-level measures for our analysis: school value-added, which is used as our proxy of school quality, and our measure of high exposure to the policy, which is used to distinguish schools in our treatment and control groups. The following sections explain how those measures were created.

III.A School Value-Added Estimates

School value-added (VA) is a measure of a school's contribution in a given year to students' test scores. We use it as our proxy for school quality, with the assumption that this measure captures how much a school increases students' achievement, controlling for all other relevant variables. This measure of school quality is meant to capture schools' inputs such as teacher quality, infrastructure, school environment, and any other school-specific characteristic that improves student achievement, measured as the average test score.

[17] This restriction necessarily means the set of schools in our sample is positively selected. We discuss entry into and exit the educational market in Appendix Section A.2

To calculate school VA, we run the following OLS regression:[18]

$$testscore_{ist} = \alpha + \gamma_g testscore_{ist-1} + \lambda_g testscore^2_{ist-1} + \mathbf{X}_i'\delta + \beta_{st} + \epsilon_{ist} \tag{1.1}$$

where $testscore_{ist}$ is the test score for a student i, at school s in year t. Students in the third through eighth grade take both a math and an English language arts exam each year; thus, we have school VA estimates for each subject as well as for the average of both scores. These scores are standardized within grade and year so that estimates can be interpreted as standard deviations. $testscore_{ist-1}$ is the student's test score from the previous academic year and is constructed in the same manner as $testscore_{ist}$. In this specification, we cannot include third graders as they do not have a previous test score. γ_g and λ_g are grade-specific coefficients on lagged test scores and lagged test scores squared. $\mathbf{X}_i$ contains several indicators for student demographics including female, Black, Hispanic, Asian, two or more races, subsidized lunch, special education, Section 504, and testing accommodations. Our school value-added measure comes from the school-year fixed effects, β_{st}. The choice of the specification is motivated by that used in Chetty et al. [2014) to measure teacher value-added. Like Chetty et al. [2014), we control for grade-specific effects of lagged test scores to account for selection into particular schools. We also show in Appendix Table A.1 that our results are robust to the use of an empirical Bayes shrinkage procedure in our value-added estimations [Kane and Staiger, 2008).

Figure 1.1 depicts the density plots of our school value-added estimates for both the public and choice schools in our sample. Panel A shows the different distributions in the years before the policy was implemented, while Panel B plots our estimates in the years after expansion. For each panel, we report the p-value for the Kolmogorov-Smirnov equality-of-distributions test. In the years before the policy, choice schools outperform public schools, and this is confirmed with Kolmogorov-Smirnov test. After expansion, we cannot statistically distinguish between the dis-

[18] In Appendix Table A.2 we show that our results are robust to different specifications of this regression. Specifically, we re-run our difference-in-differences where school value-added is estimated using Equation (1) without any demographic controls or prior test scores (Column 2), only including demographic characteristics (Column 3), and including demographic characteristics and linearly controlling for prior test scores (Column 4).

tribution of value-added for public and choice schools. The following sections of this paper will separately analyze the changes in public and choice school quality.

III.B Construction of Exposure Measure

Our main measure for each school's exposure to the voucher policy relies on the radial distance between the physical address of each of the public schools in the sample and all of the eventual choice schools in Indiana. A public school is considered to face high exposure to the voucher policy if the nearest eventual choice school is within five miles of its location.[19] We find that around half of the public schools in the sample have at least one nearby choice school.[20] Public schools whose nearest choice competitor is outside the five-mile radius comprise our control group. Nearly all choice schools (over 98%) are located within five miles of a public school; therefore, we distinguish between high-exposure and control choice schools by where they fall in the distribution of the number of public schools within five miles. High-exposure choice schools are those in the top tercile of this distribution, with the control group then making up the bottom two-thirds.

Table 1.1 reports summary statistics for the high-exposure and control public schools in the academic year before the policy intervention. Column (1) presents the sample means of the variables for high-exposure school; Column (2) presents those same means for the schools in the control group; and Column (3) presents the results of a t-test for the difference between the two groups. High-exposure schools are different from those in the control group on several dimensions. High-exposure public schools were larger, with an average of 262 students taking the state exam versus 217 in control schools. They also had a smaller share of their students identified as White, 65% versus 91%; had a larger share of students identified as Black; 16% versus 2%; and had a larger share of students qualify for subsidized lunches, 55% versus 42%.[21]

[19]The results are robust to this definition of having a competitor. Appendix Table A.3 presents our results using 3, 5, 8, 10 and 15 miles as the required distance.

[20]Appendix Figure A.3 shows the distribution of the distance between each public school in our sample and their nearest choice school

[21]The differences in the demographic make-up of the two groups of schools are at least partly explained by their locations within the state. Appendix Figure A.2 shows the location of each public school in the sample. Public schools with a nearby choice competitor are often located in the most populous and urban counties in Indiana, while those in the control group are spread out across the more rural parts of the state.

These differences in demographics, however, do not translate to significant differences in our outcome measures of interest. High exposure public schools had an average overall school value-added estimate of 0.021 in the 2010-2011 academic year versus an average of 0.018 for the schools in the control group. In that same year, high exposure schools had an average school math value-added estimate of 0.025 and an average school reading value-added estimate of 0.009. Schools in the control group had an average of 0.026 and -0.002 in their school math and reading VA estimates, respectively. We find a similar pattern in the comparison between high-exposure and control choice schools, presented in Table 1.2. Importantly, our empirical strategy does not rely on the equality of the pre-policy summary statistics. Instead, identification requires that the change in outcomes for the control group are what those facing high exposure would have experienced had the policy not been put in place. We discuss this assumption in further detail in later sections.

IV Reduced-Form Empirical Strategy

To estimate the effects of introducing (and expanding) private school vouchers in Indiana we use a difference-in-differences model that relies on plausibly exogenous variation in a school's exposure to the voucher policy. We compare the change in school value-added in the years before and after the implementation of the policy in schools facing high exposure to the policy versus those in the control group. The underlying assumption in this strategy requires that, in expectation, the change in outcomes for the schools in the control group reflect what the schools facing high exposure would have experienced had the voucher policy not been implemented. While this assumption is ultimately untestable, we address this concern by reporting the results of an event-study specification that allows the effect of the voucher program to vary by years since implementation.

We implement this difference-in-differences (DID) strategy using the following regression:

$$VA_{st} = \beta_1 Post_t \cdot HighExposure_s + \sum_{t=2007}^{2018} \Psi_t(\mathbb{1}\{year = t\} * X_s^{2007}) + \alpha_s + \gamma_t + \epsilon_{st} \quad (1.2)$$

where VA_{st} is our constructed measure of value-added in school s at year t; $Post_t$ is an indicator that equals one in the years after the voucher policy was introduced; $HighExposure_s$ is an indicator that equals one if the public school is identified as having a nearby choice school; α_s is a school fixed effect that removes any time-invariant characteristics about the school that could otherwise bias our results; γ_t is a standard year fixed effect and ϵ_{st} is our idiosyncratic error term. Ψ_t captures the potentially time-varying effects of X_s^{2007}, a vector of initial school-level characteristics.[22] The parameter β_1 is the coefficient of interest and captures the average difference between the high-exposure and control schools in the years after adoption of the voucher policy relative to the years before. All standard errors allow for arbitrary correlation in errors at the school level.[23]

We visually test the validity of the common trends assumption by presenting a set of event-study results that allow the effect of adopting a voucher policy to vary by years since implementation. Specifically, we run the following regression:

$$
\begin{aligned}
Y_{st} = \sum_{l=-5, l\neq -1}^{6} \theta_l HighExp_s \cdot \mathbb{1}\{t - 2012 = l\} + \sum_{t=2007}^{2018} \eta_t(\mathbb{1}\{year = t\} * X_s^{2007}) \\
+ \pi_s + \lambda_t + \mu_{st}
\end{aligned}
\tag{1.3}
$$

where l represents the lag or lead of interest, and 2012 is the year of adoption. Since we omit the year before the adoption of the policy, each θ_l captures the effect of being a school facing high exposure relative to the year before the introduction of the voucher program.

Our estimation strategy bypasses the concerns present in the current difference-in-differences literature because (1) we do not exploit variation across groups treated at different times [Goodman-Bacon, 2021); (2) our main specification relies on a binary measure of treatment [Callaway et al., 2021); and (3) we do not use time-varying covariates in any of our analyses [Caetano et al., 2022). Furthermore, adding school-level, time-varying characteristics may be inappropriate in this con-

[22]In Appendix Table A.4 and A.5, we show our results are robust to the exclusion of baseline covariates and the use of a continuous measure of the number of nearby choice schools, respectively.

[23]One may be concerned that our standard errors are incorrect in this specification as we are using an estimated variable as our outcome variable of interest. To address this issue, we perform a bootstrapping procedure as described in Appendix Section A.3. We find that our estimates are more precise under this procedure, most likely because clustering at the school level significantly increases our standard errors. We, therefore, continue with our preferred specification.

text. Characteristics such as the share of students eligible for subsidized lunches may change in the post-period as a direct result of the policy; hence their inclusion in our specifications would bias our results.

V Effects of ICSP on Public School Quality

We begin by describing the estimated effects of the Indiana voucher program on public schools with a nearby choice competitor. Figure 1.2 depicts the density plots of our school VA estimates for the public schools in our sample across two periods, pre-2011, and post-2013 to align with the policy time horizon. Panel A shows the kernel density plots for the high-exposure public schools, and Panel B plots the data for the public schools in our control group. For schools facing high exposure, the distribution of school value-added after voucher adoption is clearly to the right of the distribution before the policy was implemented. For schools in the control group, the distributions are statistically indistinguishable.[24] The p-values for the Kolmogorov-Smirnov equality-of-distributions test confirm this finding. While not a formal difference-in-difference design, Figure 1.2 provides a visual preview of our findings.

The results of our main analysis are reported in Table 1.3. Each cell in the first row of the table represents the coefficient on the $Post_t \cdot HighExp_s$ interaction for separate regressions. In the second row, we include an interaction term to indicate whether a school with a nearby choice competitor also had an above-median share of its students who qualified for subsidized lunches in the year before the voucher program was introduced.[25] Each column shows the results for an individual outcome of interest. Columns (1) and (2) present the results on overall school VA; columns (3) and (4) present the results on school math VA; and columns (5) and (6) present the results on school reading VA.

Schools with a choice competitor within five miles saw an overall increase in their School VA

[24] We support this claim by running the difference-in-differences specification on the set of control schools (arbitrarily identifying high exposure as a choice school within 8 miles of its location) and find no changes in school quality. The results are shown in Appendix Table A.8.

[25] This interaction term isolates the impact of the voucher program on the set of schools facing the highest threat of competition. They are located near at least one choice school and have a high share of students that would actually qualify for the voucher.

by 0.023 of a standard deviation in the post-policy period. The estimates in column (2) show that this result is driven by schools having a nearby competitor and an above-median share of students who qualified for subsidized lunch in the year before voucher adoption. Specifically, this set of schools saw an increase in overall school VA of 0.039 (0.030 + 0.009) of a standard deviation following voucher implementation. When we look at the results for math and reading separately, we find that a similar pattern holds. On average, schools with a nearby choice competitor saw an increase in their school math VA by 0.03 s.d. and an increase in their school reading VA by 0.013 s.d. in the post-policy period. When we include the interaction terms in columns (4) and (6), the results show that schools with a high share of students who qualify for subsidized lunch saw even larger increases: 0.047 of a standard deviation in school math VA and 0.028 in school reading VA.[26]

The result that the voucher program induced an increase in school quality experienced by public school students is significant. Increased schooling quality is associated with better educational outcomes including increases in the likelihood of college attainment [Deming et al., 2014) and increases in the likelihood of attending a college with a larger share of STEM degrees [Shi, 2020). Therefore, our results not only suggest that voucher programs at scale can induce responses by schools, but they can do so in such a way that meaningfully changes the educational outcomes of students not participating in the program.

Our findings also complement the results found in Waddington and Berends [2018) that explore the effect of ICSP on the students that use the voucher. The authors use a matched difference-in-differences design to compare students that used a voucher to those that qualified and remained in public schools. They find that voucher students see significant declines in math scores and no changes in reading scores following the switch to a choice school. While the authors do not speculate on the mechanisms that could explain their results, our estimates suggest that the improvements in public school quality, particularly in math, can at least partially explain the declines they report.

[26]We also show that our results are stronger when we eliminate public schools that have a choice school within 3-8 miles of their location (Appendix Table A.9).

ICSP was adopted and expanded in two separate academic years (2011-2012 and 2013-2014, respectively). One might then wonder if the two events had differential impacts on public-school quality. We answer this question using our event-study specification. The results of Equation (3) allow us to look at the effect (relative to the year before adoption) of facing choice school competition in each year of the sample rather than averaging across the entire post-policy period. We can then compare the results at the year of expansion to that of the year of adoption to get a sense of which event is driving the results. Figure 1.3 plots the results of Equation (3) for each school quality measure of interest. Years 0 and 2 indicate the years of adoption and expansion, respectively. This figure shows a small and statistically significant jump in school math VA in the year of adoption of ICSP; however, the effects are largest across our measures of interest in the year after ICSP's expansion. Interestingly, these results suggest that despite facing the threat of losing students as the program is adopted, public schools do not seem to respond until a much larger percentage of the student body qualifies to participate. This finding suggests that we may not expect voucher programs to have these indirect effects on educational outcomes until these programs are brought to scale.

While these estimates are modest in magnitude, they are statistically significant and indicate a positive relationship between the threat of choice school competition and public school quality. We cannot make exact comparisons between our results and that of the extant literature as we are analyzing school VA rather than pure student test scores; however, our results are similar to the aggregated school-by-year estimates shown in Figlio and Hart [2014). We have also estimated models at the student-school-year level and continue to see positive and statistically significant results on the effect of the threat of choice school competition on public school performance. These models are presented in Appendix Table A.6 and show that our results are similar in size to those found in the first few years after the Florida voucher program was adopted [Figlio et al., 2020).

V.A Heterogeneity by School Attributes

We have found consistent evidence of modest improvements in school VA when comparing public schools facing the threat of choice school competition to those in the control group. However, these average estimates across all public schools facing competition could differ across various subgroups. Therefore, we disaggregated the results by the following baseline characteristics: enrollment, overall school VA and median income of the census block group where the school is located. We calculated these estimates by introducing interactions of the school subgroup with the $Post_t \cdot HighExp_s$ indicator in Equation (2).[27]

Table 1.8 displays the results of our heterogeneity analysis by school subgroup for overall school VA, school math VA, and school reading VA, respectively. Panel A displays the differences in outcomes for public schools with above- and below-median enrollment for the 2006-2007 academic year. Across all of the columns, the estimate on the interaction term with above-median baseline enrollment is statistically insignificant. This result implies that public schools see similar improvements in quality when facing the potential threat of competition regardless of whether they have relatively small or large baseline enrollment.

In Panel B, we examine the differences in outcomes for public schools with above- or below-median overall school value-added for the 2006-2007 AY. Across all outcome variables of interest, the estimate on the interaction term with above-median baseline school VA is negative, statistically significant, and almost equal in magnitude to the overall estimate on the $Post_t \cdot HighExp_s$ indicator. These findings imply that the changes we see in school quality are driven by the schools that face potential competition and were originally low-performing. In fact, high-exposure schools with above-median baseline school value-added see small or no changes in the outcomes of interest when compared to the control group. The increase in school quality for low-performing schools, coupled with the null results for high-performing schools, suggests that the gap in public-school

[27] We have also considered heterogeneity by initial levels of suspension/expulsions. This analysis addresses a different type of threat public schools could face. Specifically, families may have a desire to leave public schools that they deem unsafe. We do not find any differential effects for public schools that had an above median percentage of their students ever being suspended or expelled. Results are available upon request.

quality is closing as a result of the program.[28]

Panel C reports the effects on quality by the mean income of the census block group where the public school is located. This specification allows us to capture any differences in the results between public schools located in relatively rich versus poor neighborhoods. Similar to the results in Panel A, the estimate on the interaction term with above-median neighborhood income is statistically insignificant across all outcomes of interest. These findings imply that schools see similar improvements in quality when facing the potential threat of competition regardless of whether they are located in a relatively poor or rich neighborhoods.

We also explore possible heterogeneity by financial incentive. As shown in Figlio and Hart [2014), not all public schools face the same incentives to respond to the implementation of a voucher program. Specifically, public schools on the margin of receiving federal Title I aid may experience a larger reduction in resources as a consequence of losing students to private schools. We, therefore, explore whether high-exposure public schools with Title I funding drive our results. Panel A of Appendix Table A.7 reports the differences in outcomes for public schools with and without a Title I program in year before ICSP was adopted. Panel B of Appendix Table A.7 includes an interaction term that allows us to identify the differential impact of ICSP on high-exposure public schools that just qualified for Title I funding.[29] Overall, we do not find evidence that public schools facing greater financial pressure respond more to the program.

V.B Potential Mechanisms

V.B.1 Changes in School Inputs

Given the improvements we find in pubic-school quality, we next examine changes in schools inputs that might lead to increases in school quality. In particular, we combine information from the Common Core of Data on Indiana public schools from the National Center of Education Statistics

[28]One may be concerned that these results are driven by families wishing to leave low performing public schools; however, as shown in Appendix Figures A.6 and A.7 there does not seem to be differential student sorting on ability across these two types of public schools when comparing either FRPL (A.7) or non-FRPL students (A.7)

[29]Title I funding is allocated based on where a school ranks within their districts' with respect to the share of low-income students they serve. In Indiana, schools that meet or exceed the district's poverty average are eligible to receiving funding. We define "just qualifying" for Title I as being within 5 percentage points above that cutoff for eligibility.

with available teacher data in the IDOE-CREO database to explore changes in student-teacher ratios, the number of teachers with a high-quality (HQ) certification,[30] number of teachers with a graduate degree and teachers' average years of experience. Unfortunately, the information on teachers is only available from the 2010-2011 through 2017-2018 academic years, which limits our sample to include only one year of pre-policy data.

Figure 1.5 separately plots the average of each of these school inputs across the available years of data for high-exposure and control public schools. We do not find strong evidence that high-exposure public schools saw meaningful changes in student-teacher ratios or the average years of experience of their teachers when compared to the control group. However, Panel B shows that while both high exposure and control public schools added around 2 additional HQ certified teachers (either through hiring or certification) in the year ICSP was adopted, control public schools did not retain them. By the end of the sample period, control public schools had returned to their initial levels of HQ certified teachers. Furthermore, Panel C, shows that while both high-exposure and control public schools see declines in the average number of teachers with a graduate degree, control public schools witness faster declines over the sample period.

We confirm these patterns in the data with the results from our difference-in-differences specification. In Table 1.4, we report the results of Equation (3) using school inputs as our outcome measures of interest. We find that relative to the year before ICSP was adopted, high-exposure public schools saw increases of around 0.7 teachers with a graduate degree and 1.5 teachers with a HQ certification when compared to the control group.[31]. These changes in average teacher characteristics are significant. While the previous literature on the effects of advanced degrees on student outcomes is mixed, recent work shows that subject-specific teacher credentials (such as a high-quality certification) are associated with stronger student achievement [Strøm and Falch, 2020).

We also examine the impact of ICSP on students' non-cognitive skill formation in public

[30] High-Quality certification is determined by standards set by No Child Left Behind. States are allowed to add their own requirements. For the state of Indiana, HQ certification requires passing an additional exam to indicate proficiency in a certain subject.

[31] Our results differ from those in Figlio and Hart [2014) The authors find that schools faced with greater competition shift their teacher workforce to include less- qualified teachers. Unfortunately, we lack the detailed data on school practices to fully disentangle different school responses under each of these reforms.

schools. Table 1.5 reports the results of Equation (3) where the outcomes of interest are school-level measures of attendance and disciplinary infractions. These two measures have been cited as important indicators for changes in behavior [Imberman, 2011b). After the implementation of ICSP, public schools facing the threat of choice school competition saw increases in attendance and decreases in suspensions/expulsions. Specifically, high-exposure public schools saw increases in attendance of 0.3 percentage points (p.p.), or about half a day, compared to those in the control group. The estimate in column (3) suggests that high-exposure public schools also saw a reduction of 0.5 p.p in expulsions and suspensions, with the caveat that this estimate is statistically insignificant. Attendance is cited as important determinant of student outcomes including test scores [Goodman, 2014, Fitzpatrick et al., 2011, Gottfried, 2009) and high school graduation [Liu et al., 2021). Using the estimates in Goodman [2014), we can do a back-of-the-envelope calculation that reveals that the increase in attendance by half a day, induced by ICSP, can translate into around a 0.025 s.d. deviation increase in test scores.[32] Overall, we take these results as evidence that in response ICSP, schools are increasing quality such that we see improvements on both cognitive and non-cognitive dimensions.

V.B.2 Changes in School Financial Resources

ICSP could further have a direct effect on public schools' ability to improve school quality through changes in financial resources. Opponents of school choice policies argue that these programs drain public school finances through direct cuts in state funding [Strauss, 2017). Moreover, losing students eligible for subsidized lunches could result in further resource reductions if schools rely on Title I funding. By contrast, per-pupil revenue may increase in public schools if total federal and local funding remain unchanged.[33] If the latter is the case in Indiana, increases in available school funds could contribute to our results.[34].

[32] One does need to keep in mind that the estimate from Goodman [2014) has a very specific interpretation, as it is identified off of missed classes due to snowfall, that may not translate to our context.

[33] DeAngelis and Trivitt [2016) show that if Louisiana Scholarship Program was eliminated only 2 to 7 out of 69 school districts would see an increase in financial resources.

[34] There still remains some debate on whether increases in school spending improve educational outcomes [Jackson, 2020). One direct way increased school per-pupil expenditure could directly influence our results is if schools used the extra funds to hire or convert high quality teachers. This is left as an open question as we do not have the data to

However, school funding in Indiana heavily relies on state rather than local sources. The state currently ranks 40$^{\text{th}}$ in the percent of public school funding coming from local revenues (just below 30%) [U.S Census Bureau, 2021). Furthermore, the state has provided 100 percent of funds available to support education-related operating costs since 2009. Local funds are used to support other expenses including transportation, capital projects, and debt services [Chu, 2019). This reliance on state-funding suggests that Indiana public schools are susceptible to reductions in revenues as students use the voucher. Anecdotal evidence from statements made by public school boards echo this concern [Gore et al., 2011). Unfortunately, school-level finance data is not available for a majority of our sample period; therefore, we cannot formally test whether changes school funding can explain our results. Future research will explore this question at a greater length.

V.C Student Sorting

The results from the previous section suggest that ICSP implementation improved public-school quality; however, it is necessary to distinguish between whether the results we find are due to actual changes made by schools or are driven by the composition of students that remain in the public schools. In this section, we present evidence suggesting that the sorting of students, while apparent, cannot explain all of the gains in school value-added we report.

We first investigate this issue by documenting any changes in the demographic composition of students in high-exposure public schools after the implementation of the program. Table 1.6 reports the results of Equation (3) where the outcomes of interest are school-level measures of demographic variables (Share Female, Share White, Share Black, etc.). After the implementation of ICSP, public schools facing the threat of choice school competition saw statistically insignificant changes of -0.19 p.p in the share of students that are female, 0.27 p.p in the share of students that are Black, and 0.38 p.p in the share of students qualifying for subsidized lunch when compared to the control group. However, as shown in columns (2) and (4), high exposure public schools saw a statistically significant decrease of -2.72 p.p in the share of White students and an increase of 2.27 p.p in the share of Hispanic students.

---

test this theory.

We next address the concern of student sorting on ability. Figure 1.4 shows the density plots of standardized test scores for students who eventually use a voucher and those students who remain in the public-school system despite qualifying to participate in the program. Specifically, the figure plots the standardized test scores in the years before the program was adopted. We find that eventual voucher students slightly outperformed those remaining at the public schools. This finding suggests that ICSP did induce some “cream-skimming", which has been a major criticism of voucher policies. However, this type of sorting on ability works against the theory that the students leaving the public school system would artificially increase average test scores.[35]

Overall, we take these results as evidence that the demographics of students are changing with the implementation of the voucher program. To understand to what extent these changes in demographics drive our results, we perform an exercise with predicted school value-added. Specifically, we begin by estimating the following model:

$$VA_{s,2007} = \sigma X_s^{2007} + \epsilon_s \tag{1.4}$$

where $VA_{s,2007}$ is our estimated school value-added in 2007 (our “base” year), and X_s^{2007} includes all of the school characteristics we observe and their pairwise interactions in that same year. We use the coefficients from this fully interacted model to predict value-added for each school in all years of the sample. We then use these predicted value-added measure to run the following difference-in-differences specification:

$$\hat{VA}_{st} = \beta_1 Post_t \cdot HighExposure_s + \alpha_s + \gamma_t + \epsilon_{st} \tag{1.5}$$

If changes in observable school characteristics are driving our school quality results, we would expect differential changes in the predicted value-added measures following the implementation

[35]We do not have information on whether students remaining in the public school qualify for a 50% voucher; hence the comparison made in Figure 1.4 also compares 50% voucher students to FRPL students. Appendix Figure A.5 shows the direct comparisons of eventual choice students versus those remaining in the public schools system for both FRPL and non-FRPL groups in Panels A and B respectively. We find almost no sorting on ability when comparing FRPL students and a slight negative selection for non-FRPL students. However one must consider that the non-FRPL comparisons also include high-income students that do not qualify for a voucher.

of the voucher policy. Table 1.7 reports the results of this exercise. We find no evidence that high-exposure public schools were predicted to improve their quality based on the change in composition of their students. In fact, we find that based solely on changes in observable characteristics, high-exposure public schools were predicted to see declines in overall and math value-added. We take this result as strong evidence that it is changes made by schools that drive the improvements in quality we see. We also recognize that this exercise can only speak to how changes in observable school characteristics may have affected our school-quality results. The concern remains that non-random student sorting on unobservable characteristics is driving our results.

We can mitigate some concerns of non-random sorting on unobservable characteristics by highlighting the strength of our value-added estimation strategy. In Equation 1.1, we control for lagged test-scores. Assuming that prior test scores fully proxy for those inputs that affect a student's achievement prior to using the voucher and that those inputs correlated with a student's likelihood of using a voucher, we address the concerns for this type of sorting. This is a strong assumption, however, it is standard in the school value-added literature.

V.D Threats to Validity

The previous section shows that ICSP implementation is associated with increased school value-added estimates for public schools with a nearby choice school. There remain, however, several potential threats to validity that should be addressed. Specifically, (1) the impact of the voucher policy on high exposure public schools may be driven by differential trends in school value-added across the two groups of schools before program implementation, (2) the results may be sensitive to the exclusion of particular districts that house a large proportion of the students in the state, and (3) there are other policy innovations besides the voucher program that may be driving the results.

To ensure that the findings are not driven by differential trends between the schools facing high exposure to the voucher policy and the control group, Figure 1.3 plots the event-study results of Equation (3) for each school-quality measure of interest. This analysis gives a sense of when school VA patterns changed and if preexisting trends are driving the results. The coefficients are plotted with 95 percent confidence intervals; the omitted category is the schools in the year prior to the program implementation. The expansion of the voucher program is highlighted at Year 2, which corresponds to the 2013-2014 academic year. Prior to implementation, high-exposure public schools and the control group appear to have similar trends in school value-added, shown by the relatively flat differences between the two groups.[36]. In all years before implementation, the 95 percent confidence interval contains zero, which means that in those years, the difference between high-exposure and control groups cannot be distinguished from the value in the year before implementation.[37]

The second concern is that the results are sensitive to the exclusion of particular school districts. We, therefore, estimate the main analysis in Table 1.3 excluding Marion County, the largest county in the state and the home of Indianapolis. We find consistent evidence that, regardless of dropping Marion County, the signs and general significance levels of the interaction term of interest hold as shown in Appendix Table A.10. Appendix Table A.11 shows that when we drop any of the 92 Indiana counties, our results remain similar to the full-state analysis. Therefore, it is difficult to believe that some combination of specific counties are driving the general direction of our results.

Another concern is that other policy interventions beyond the voucher program are driving the results. To address this issue we use year fixed effects in each of our specifications to capture shocks common to both the treatment and control groups. Unaccounted for shocks could still exist, but those shocks would have had to elicit disproportionate reactions from schools with a nearby choice competitor to account for our results. A particular concern is that in 2011 the implementation of the Teacher Evaluations and Licensing Act and the introduction of Indiana's A-F school grading system may have affected school quality. However, since the quality of schools in the high exposure and control groups were statistically indistinguishable in 2010, it is unlikely

[36]We further show the robustness of our results using placebo treatment years. Appendix Figure A.4 shows the results when we assign the adoption of ICSP to be two years prior to the actual The figure shows that school quality only improved following the years of actual adoption and expansions (As indicated by the red and blue dashed lines, respectively).

[37]Appendix Figure A.4 shows the results of our event-study specification only including those public schools that had an above median share of FRPL students in 2010. We include this specification because this is the group of public schools that drive our main results in Table 1.3.

that either of these reforms differentially impacted the two sets of schools. Moreover, it is not clear whether schools felt increased pressure to improve quality as a result of these accountability programs. Prior to the adoption of these specific measures, schools and teachers were held to other accountability metrics. Furthermore, in the 2013-2014 academic year, less than 0.5 percent of teachers were cited as “ineffective" and only 4 percent of public elementary and middle schools were given an “F" grade [Indiana Department of Education, 2014a,b).

VI Effects of ICSP on Choice School Quality

Our results thus far have been centered on public schools' responses to the implementation of ICSP. We next assess whether participating private schools also saw changes in school quality as a result of the program. This investigation is necessarily more speculative than our analysis of public schools due to data constraints.[38] However, in this section we present evidence that choice schools are reducing quality after the adoption of ICSP.

We first investigate choice schools' response to the adoption of the voucher program by plotting the averages of our school value-added measures for each year in the sample. Figure 1.6 plots these averages for our measures of school quality from 2007 through 2018. In the first year of the program, there is an immediate drop in average quality on all dimensions. This drop is most apparent for math value-added, but by the following year, the average reading value-added for choice schools saw a similar decline. These school quality measures, while steadily increasing after 2013, remain below the pre-period levels until 2016 for reading and throughout the sample period for math. While we do not assert any causal claims from this figure, it does suggest that choice schools saw a decline in quality following the implementation of the program.

Ideally, we would be able to examine choice schools' responses to the implementation of ICSP by comparing them to the set of private schools that never accepted voucher students. Unfortunately, we do not have data on a large percentage of non-choice private schools. We can, however, compare choice schools that pull students from a large pool of public schools to those with fewer

[38] Specifically, we are unable to compare choice schools to non-choice private schools since non-choice private schools often do not use the ISTEP+ exam, and we are unable to leverage variation in when a choice school starts accepting voucher students as a large percentage adopt in the first year of the program.

public schools in the area. We can then assess whether choice schools responded differently to the voucher program based on the potential number of students they could receive.[39] High exposure is now defined as being in the top tercile of the distribution of the number of public schools within a five-mile radius.

Figure 1.7 shows the density plots of our school VA estimates for these groups of choice schools across two time periods: Pre-2011 and Post-2013 to align with the program's adoption and expansion. Panel A shows the kernel density plots for high exposure choice schools and Panel B plots the data for those choice schools in the control group. Both groups witness a leftward shift in the distribution of overall school value-added following the expansion of ICSP, suggesting that ICSP may not have elicited differential responses across our measure of exposure. Table 1.9 formalizes this comparison using our difference-in-differences specification (similar to Equation (3)). Columns (1), (2), and (3) present the results on overall school value-added, school math value-added, and school reading value-added, respectively. After the implementation of ICSP, treated choice schools saw statistically insignificant decreases of around 0.01 s.d. across each of our measures of school quality when compared to the control group. This exercise ultimately cannot explain the large drops in school quality seen in Figure 1.6 but suggest that choice schools with a larger pool of students to pull from saw larger drops in school quality.

To understand what is driving the declines in quality we find, we use data from the Private School Universe Survey to examine changes in choice-school inputs. Specifically, we have information on the number teachers, student-teacher ratios, and the time spent in school (in hours) every other year from 2006 until 2018. Figure 1.8 plots the averages of these inputs separately for high-exposure and control choice schools. We find that following the adoption of ICSP, there is evidence that high exposure choice schools experienced an increase in their student-teacher ratios. Panel C shows that while both high exposure and control choice schools increased their average in instructional time, control choice schools saw a more significant rise. We confirm these find-

[39]We also show results in Appendix Table A.12 that alter the definition of high exposure for choice schools. Rather than distinguishing between treatment and control based on the distribution of the number of public schools within five miles, we split choice schools by the percentage of public school students that would qualify for the voucher in the schools within five miles of their location. We find similar results under this specification

ings with the results from our difference-in-differences specification shown in Table 1.10. We find evidence that following the adoption of ICSP, high exposure choice schools saw a statistically significant increase of 0.93 in their student-teacher ratio (off a base mean of 14.24) compared to the control group[40] with the results on the number of teachers and instructional time being statistically insignificant. We, therefore, conclude that once we include baseline controls, the differences in these inputs across high-exposure and control choice schools are no longer apparent.

Evidence from Project STAR reveals that changes in student-teacher ratios can have a significant impact on student outcomes, including test scores [Krueger, 1999), high school graduation [Finn et al., 2005), college entrance exam taking [Krueger and Whitmore, 2001), college matriculation [Chetty et al., 2011), criminal activity, and teen birth rates [Schanzenbach, 2006). Therefore, our result that students in high-exposure choice schools experience increases in their class sizes further shows that voucher programs at scale can have important impacts on the educational outcomes of students that do not participate in the program.

VII Quantifying the Incentives to Change Quality

Our results so far demonstrate that public schools responded to the implementation of the voucher policy by increasing school quality, while participating private schools reduced quality. However, there remains the question of why schools would alter their quality in response to the voucher program. Prior work argues that potential changes in enrollment or market share serve as a main mechanism [Epple et al., 2017). In this section, we test this commonly hypothesized mechanism and estimate its relationship to changes in schools' incentive to provide quality.

Specifically, we estimate a structural model of household demand for schools with a partial model of how schools choose quality and tuition (public schools do not choose tuition). Our model of demand incorporates rich amounts of heterogeneity in how households trade-off tuition, quality, and distance which allows us to fit flexible substitution patterns from the data. Our model

[40] Our results are similar in magnitude (around a 7% increase versus 9% from the authors results) to those found in Rinz [2015) that examines changes in private school inputs following the adoption of voucher programs throughout the 2000s. His analysis includes both traditional voucher programs and large scale tax credit programs, which shows that these two variations of voucher programs may have similar impacts on private school responses.

of school incentives (particularly for private schools) is similar in spirit to Neilson [2021). In estimation, we exploit variation in the program's eligibility, voucher amounts, and spatial distribution of students and schools for identification. A full outline of the model specification and estimation are available in Appendix Sections A.4 and A.5. We quantify the potential changes in enrollment (market share) for both public and private schools by using our demand estimates to simulate how household choices would change under counterfactual voucher scenarios holding all else constant. We alter the design of the voucher program along two dimensions. First, we change the maximum voucher amount currently eligible students are able to receive. Second, we keep the voucher amount constant while changing the income threshold for eligibility. For high-income households, eligibility amounts to receiving a 50% voucher.[41] As we alter the program's design we estimate the average number of students attending public and choice schools holding fixed other variables in the market including school quality, tuition and distance. We then interpret the changes in market share as the potential loss (or gain) faced by the schools absent any other changes.

We then map this enrollment threat into changes in incentives to provide quality by modeling public and private schools. Importantly, we do not take a direct stance on the objective function of public schools. Instead, we rely on changes in the quality elasticity of demand, while we cannot directly interpret the magnitudes when examining changes, understanding its direction provides a clear picture of how incentives change for public schools regardless of their specific underlying objective function. For choice schools, we are able to quantify changes in their market power by estimating their ability to mark down quality (value-added) below competitive levels.[42]

Counterfactual I: Changing Voucher Amounts

Figure 1.9 shows how average enrollment at public and private schools changes as we vary the voucher amount from 0% to 200% of the current number. For each of our 15 scenarios, we alter the voucher amount based on a multiplier. For example, a multiplier of zero suggests there is no

[41] Each of these counterfactual voucher schedules are motivated by current proposals of expansions of ICSP from the Indiana State Legislature.

[42] As we outline in Appendix Sections A.4 and A.5, the measures we propose to estimate the incentive to provide quality capture how sensitive demand is to changes in quality. The quality markdown, however, is more precisely related to the inverse quality elasticity of demand.

voucher program, and a multiplier of two suggests students are eligible to up to double the current voucher amount. These counterfactual amounts range from $0 to $\sim \$6,000$ and $\sim \$10,000$ for students eligible for the 50% and 90% vouchers,[43]respectively[44]. Importantly, we do not make changes to the eligibility thresholds under this counterfactual.

We find that public schools face a large threat in market share (enrollment) once the voucher amount is large enough to induce households to switch to private schooling. Specifically, Figure 1.9a shows that the average public school would not see any changes in market share until the voucher reaches 60% of the current amount. As the voucher amount further expands, we begin to see a significant threat to enrollment with the average public school seeing a near a 25% decline in their market share absent any changes in school quality. This potential market share drop is mainly driven by students who qualify for a 90% voucher. Students eligible for a 50% voucher drive the second, smaller drop in market share that occurs at 120% of the current voucher amount. Figure 1.9b shows the analogous rise in market share that occurs for private schools at each of those jump points.

Figure 1.10a shows how these changes in market share map into the incentives to provide quality. We find that, all else constant, the quality elasticity of demand increases with the potential outflows of students from public to private schools. This result suggests that public schools' incentives to improve quality are increasing as the voucher amount increases. However, there is a flattening of the quality elasticity of demand after the voucher amount exceeds 150% of the current voucher amount, suggesting the incentive to improve quality is increasing only up to a certain point.

Figure 1.10b shows that private schools have an incentive to mark down quality as voucher amounts increase. The overall pattern matches that of market share. There is a slight decrease in the magnitude of the quality markdown as the voucher exceeds 150% of the current voucher amount which we argue is driven by 50% voucher students switching to private schools. Our

[43]These amounts are based on funding for the 2017-2018 AY

[44]No matter how the size of the voucher changes, students always receive the lesser of tuition and the maximum voucher amount.

demand estimates suggest that 50% voucher students are more sensitive to changes in quality than their lower-income counterparts; therefore, private schools are not able to markdown quality as much if they wish to attract this group of students.

Counterfactual II: Changing Voucher Eligibility

Figure 1.11 shows how average market share at public and private schools changes as we vary the income threshold necessary for eligibility. We set the voucher amount households receive equal to that of the current program. We do not change the current program's cutoff to receive a 90%, but instead increase the cutoff to receive a 50% voucher. Therefore, as we increase the income eligibility threshold beyond the current program, newly eligible households would receive a 50% voucher. Households whose income falls below the current program's 90% cutoff will always receive for a 90% voucher in scenarios they meet the counterfactual income threshold. These counterfactual income thresholds range from household incomes of $0 (the voucher program does not exist) to $180,000.

We find that, all else constant, average public school market share decreases (mostly) monotonically as income eligibility increases.[45] The slope of this curve is larger in magnitude when lower-income households become eligible, suggesting that the threat to public schools increases at a decreasing rate as the program is expanded. Figure 1.12b shows that private schools face an analogous story in the opposite direction.

Figure 1.12 shows how these changes in market share map into the incentives to provide quality. For public schools, we find that the quality elasticity of demand increases with the potential outflows of students from public to private schools. This result suggests that public school incentives to improve quality are increasing as the income eligibility threshold increases. However, as the highest-income students become eligible the quality elasticity falls to a level just above the case where a voucher policy does not exist. We argue this is due to three factors: (1) Figure 1.11a shows that public schools lose fewer students as the income thresholds increases, suggest-

[45]The step- function-like nature of the plots is a result of the demand specification, and more importantly, the voucher program. The voucher amount for 50% voucher students is not large enough to induce significant changes in enrollment even with expanded eligibility for those students. See Appendix Section A.4 for further model details.

ing there are few high-income individuals that attend public schools,[46] (2) our demand estimates show that high-income students value quality more than their lower-income counterparts and (3) public schools have lower levels of quality than private schools. Together, increasing the income eligibility threshold all but ensures that the few high-income students that initially attend a public school will leave, since we hold school quality constant. Public schools then see an overall decline in their incentive to provide quality as the remaining students value quality less.

Figure 1.12b shows that private schools are able to mark down quality as more students become eligible to receive a voucher. However, the relationship is non-linear. Quality markdowns are increasing in magnitude until the income threshold reaches around $91,000. Beyond that point, private schools' markdowns are slightly smaller. This result is a direct consequence of the fact that higher-income students are more likely to already be attending a private school. Therefore at higher income eligibility thresholds, private schools must compete against each other to obtain new students. The changing nature of competitions faced by private schools shifts the incentive to provide quality up.

Overall, these counterfactuals provide two main insights. First, public schools are in fact threatened by the possibility of students leaving and this threat leads to public schools having incentives to improve school quality. However, this threat only exists when the voucher is significant enough to induce to students to possibly leave. This is a direct result of preference heterogeneity (mainly price sensitivity) of nearby households. We estimate larger quality markdowns for private schools and therefore conclude they have more local market power as vouchers are expanded. Second, there are significant nonlinearities in the incentives to provide quality as income eligibility is expanded. Our results suggest that extending voucher eligibility to higher-income individuals could erode some of the public school incentives to increase quality. However, private school quality markdowns may decrease in magnitude as voucher eligibility is expanded. We estimate that they will still be larger than in the absence of vouchers.

[46]Lower-income students are by the largest group of students enrolled in public schools.

VIII Conclusion

This paper shows that the implementation of an at-scale voucher program can lead to meaningful changes in school quality. We examined the effects of the adoption and expansion of the Indiana Choice Scholarship Program, the largest program in the United States providing private school vouchers to low and middle-income families, and found that both public and participating private schools saw changes in their school value-added.

We found that public schools facing high exposure to the voucher program experienced increases in their school quality, while choice schools witnessed declines. Our estimates were modest in magnitude; however, papers evaluating voucher policies have found relatively small effects on student outcomes ranging from -0.01s.d. to 0.11s.d [Rouse and Barrow, 2009).[47] Furthermore, Figlio et al. [2020) shows that the impact on public schools grow as voucher programs mature. We analyze the program in the first few years of its adoption, so it is possible to see stronger increases in the future.

Our results complement those found in previous work examining the effect of ICSP on students that use the voucher. Waddington and Berends [2018) shows that students participating in the program saw declines in math performance with no changes in reading. We argue that schools' responses can at least partly explain these student-level results. The results in Waddington and Berends [2018) might overstate the decline in math performance since this is the dimension that high exposure public schools saw the greatest improvements. Our results provide an example of how understanding of a program's effectiveness may change when we take into consideration the indirect effects when the policy is brought to scale.

We then use a model of household demand for schools to estimate the threat to public school enrollment and its relationship with incentives to provide quality holding tuition and quality constant in the market. Our results suggest that 1) the threat to public schools matters if the voucher amount is large enough and 2) significant non-linearities in incentives to provide quality exists when scaling this program up in terms of income eligibility. Therefore, policymakers should caution against

[47] Abdulkadiroğlu et al. [2018) and Waddington and Berends [2018) are notable exceptions.

using the successes/failures of smaller-scale voucher programs as motivation for expanding/not expanding them to more people due to the presence of these nonlinearities in incentives to provide quality. To our knowledge, these nonlinearities in incentives to provide quality have not yet been documented.

IX References

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

O. Al-Ubaydli, J. A. List, and D. L. Suskind. What can we learn from experiments? understanding the threats to the scalability of experimental results. *American Economic Review*, 107(5):282–86, 2017.

A. Böhlmark and M. Lindahl. Independent schools and long-run educational outcomes: Evidence from sweden's large-scale voucher reform. *Economica*, 82(327):508–551, 2015.

C. Caetano, B. Callaway, S. Payne, and H. S. Rodrigues. Difference in differences with time-varying covariates, 2022.

B. Callaway, A. Goodman-Bacon, and P. H. C. Sant'Anna. Difference-in-differences with a continuous treatment, 2021.

R. Chakrabarti. Can increasing private school participation and monetary loss in a voucher program affect public school performance? evidence from milwaukee. *Journal of public Economics*, 92 (5-6):1371–1393, 2008.

R. Chakrabarti. Vouchers, public school response, and the role of incentives: Evidence from florida. *Economic inquiry*, 51(1):500–526, 2013.

F. Chen and D. N. Harris. How do charter schools affect system-level test scores and graduation rates? a national analysis. *Working Paper*, 2021.

R. Chetty, J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan. How does your kindergarten classroom affect your earnings? evidence from project star. *The Quarterly journal of economics*, 126(4):1593–1660, 2011.

R. Chetty, J. N. Friedman, and J. E. Rockoff. Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American economic review*, 104(9):2593–2632, 2014.

H. Chiang. How accountability pressure on failing schools affects student achievement. *Journal of Public Economics*, 93(9-10):1045–1057, 2009.

M. M. Chingos and P. E. Peterson. Experimentally estimated impacts of school vouchers on college enrollment and degree attainment. *Journal of Public Economics*, 122:1–12, 2015.

D. Chu. Indiana's property tax, choice, and accountability reforms: Their consequences for funding and student achievement, 2019.

S. R. Cohodes and K. S. Parham. Charter schools' effectiveness, mechanisms, and competitive influence. Working Paper 28477, National Bureau of Economic Research, February 2021.

C. A. DeAngelis and J. R. Trivitt. Squeezing the public school districts: The fiscal effect of eliminating the louisiana scholarship program on state education expenditures. *University of Arkansas Working Paper Series*, 2016.

A. Deaton and N. Cartwright. Understanding and misunderstanding randomized controlled trials. *Social science & medicine*, 210:2–21, 2018.

A. Deeb. A framework for using value-added in regressions. *arXiv preprint arXiv:2109.01741*, 2021.

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

EdChoice. *The ABCs of School Choice: The Comprehensive Guide to Every Private School Choice Program in America. 2021 Edition*. EdChoice, 2021. URL `https://eric.ed.gov/?q=source%3A%22EdChoice%22&id=ED612883`. Publication Title: EdChoice.

D. Epple, R. Romano, and R. Zimmer. Charter schools: A survey of research on their characteristics and effectiveness. *Handbook of the Economics of Education*, 5:139–208, 2016.

D. Epple, R. E. Romano, and M. Urquiola. School vouchers: A survey of the economics literature. *Journal of Economic Literature*, 55(2):441–92, 2017.

D. Figlio and C. Hart. Competitive effects of means-tested school vouchers. *American Economic Journal: Applied Economics*, 6(1):133–56, 2014.

D. N. Figlio and C. E. Rouse. Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics*, 90(1-2):239–255, 2006.

D. N. Figlio, C. M. Hart, and K. Karbownik. Effects of scaling up private school choice programs on public school students. Technical report, National Bureau of Economic Research, 2020.

J. D. Finn, S. B. Gerber, and J. Boyd-Zaharias. Small classes in the early grades, academic achievement, and graduating from high school. *Journal of Educational Psychology*, 97(2):214, 2005.

M. D. Fitzpatrick, D. Grissmer, and S. Hastedt. What a difference a day makes: Estimating daily learning gains during kindergarten and first grade using a natural experiment. *Economics of Education Review*, 30(2):269–279, 2011.

M. Friedman. *Capitalism and Freedom*. Phoenix books in political science, P111. University of Chicago Press, 1962. URL `https://books.google.com/books?id=yvHrAAAAMAAJ`.

M. Gilraine, U. Petronijevic, and J. D. Singleton. Horizontal differentiation and the policy effect of charter schools. *American Economic Journal: Economic Policy*, 13(3):239–76, August 2021. doi: 10.1257/pol.20200531. URL `https://www.aeaweb.org/articles?id=10.1257/pol.20200531`.

J. Goodman. Flaking out: Student absences and snow days as disruptions of instructional time. Working Paper 20221, National Bureau of Economic Research, 2014. URL `http://www.nber.org/papers/w20221`.

A. Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277, 2021.

E. M. Gore, D. Arnold, and M. R. Zaphiriou. The foregoing resolution no. 7591 was passed by the board of school commissioners of the city of indianapolis, indiana, this 15th day of february, 2011., 2011. URL `http://go.boarddocs.com/in/indps/Board.nsf/goto?open&id=8E4QDB68C590`.

M. A. Gottfried. Excused versus unexcused: How student absences in elementary school affect academic achievement. *Educational Evaluation and Policy Analysis*, 31(4):392–415, 2009. doi: 10.3102/0162373709342467. URL `https://doi.org/10.3102/0162373709342467`.

J. P. Greene and R. H. Marsh. The effect of milwaukee's parental choice program on student achievement in milwaukee public schools. scdp comprehensive longitudinal evaluation of the milwaukee parental choice program. report# 11. *School Choice Demonstration Project*, 2009.

J. J. Heckman. Randomization and social policy evaluation revisited. Working Paper 107, National Bureau of Economic Research, July 1991. URL http://www.nber.org/papers/t0107.

W. G. Howell, P. J. Wolf, D. E. Campbell, and P. E. Peterson. School vouchers and academic performance: Results from three randomized field trials. *Journal of Policy Analysis and management*, 21(2):191–217, 2002.

C. M. Hoxby. School choice and school competition: Evidence from the united states. In *School choice and school competition: Evidence from the United States*, 2003.

C.-T. Hsieh and M. Urquiola. The effects of generalized school choice on achievement and stratification: Evidence from chile's voucher program. *Journal of public Economics*, 90(8-9):1477–1503, 2006.

D. M. Hungerman and K. Rinz. Where does voucher funding go? how large-scale subsidy programs affect private-school revenue, enrollment, and prices. *Journal of Public Economics*, 136: 62–85, 2016.

D. M. Hungerman, K. Rinz, and J. Frymark. Beyond the Classroom: The Implications of School Vouchers for Church Finances. *The Review of Economics and Statistics*, 101(4):588–601, 10 2019. ISSN 0034-6535. doi: 10.1162/rest_a_00782. URL https://doi.org/10.1162/rest_a_00782.

S. A. Imberman. The effect of charter schools on achievement and behavior of public school students. *Journal of Public Economics*, 95(7-8):850–863, 2011a.

S. A. Imberman. Achievement and Behavior in Charter Schools: Drawing a More Complete Picture. *The Review of Economics and Statistics*, 93(2):416–435, 05 2011b. ISSN 0034-6535. doi: 10.1162/REST_a_00077.

Indiana Athletic Association. By-laws & articles of incorporation, 2021. URL https://legacy.ihsaa.org/About-IHSAA/By-Laws.

Indiana Department of Education. State a-f results, 2014a. URL https://www.in.gov/

doe/it/data-center-and-miscs/.

Indiana Department of Education. Staff performance evaluation results 2013-2014, 2014b.

Indiana Department of Education. 2020-2021 choice scholarship program annual misc: Participation & payment data, 2021a.

Indiana Department of Education. 2022-2023 choice scholarship program track eligibility requirements, 2021b.

Indiana Department of Education. Indiana choice scholarship program, 2021c. URL https://www.in.gov/doe/students/indiana-choice-scholarship-program/.

C. K. Jackson. *Does school spending matter? The new literature on an old question.* American Psychological Association, 2020.

T. J. Kane and D. O. Staiger. Estimating teacher impacts on student achievement: An experimental evaluation, 2008.

A. B. Krueger. Experimental estimates of education production functions. *The quarterly journal of economics*, 114(2):497–532, 1999.

A. B. Krueger and D. M. Whitmore. The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project star. *The Economic Journal*, 111(468):1–28, 2001.

J. Lise, S. Seitz, and J. A. Smith. Equilibrium policy experiments and the evaluation of social programs, 2004.

J. Liu, M. Lee, and S. Gershenson. The short- and long-run impacts of secondary school absences. *Journal of Public Economics*, 199:104441, 2021. ISSN 0047-2727. doi: https://doi.org/10.1016/j.jpubeco.2021.104441. URL https://www.sciencedirect.com/science/article/pii/S0047272721000773.

D. P. Mayer, P. E. Peterson, D. E. Myers, C. C. Tuttle, and W. G. Howell. School choice in new york city after three years: An evaluation of the school choice scholarships program. final report. *Mathematica Policy Research Report*, 2002.

K. Muralidharan and V. Sundararaman. The aggregate effect of school choice: Evidence from a two-stage experiment in india. *The Quarterly Journal of Economics*, 130(3):1011–1066, 2015.

National Center for Education Statistics. Digest of education statistics, table 226.40, 2019. URL https://nces.ed.gov/programs/digest/d21/tables/dt21_205.80.asp. Publisher: National Center for Education Statistics.

C. Neilson. Targeted vouchers, competition among schools, and the academic achievement of poor students. *Working Paper. Yale University.*, 2021. URL http://economics.sas.upenn.edu/system/files/event_papers/Neilson_2013_JMP_current.pdf.

K. Rinz. Undone by the market? the effects of school vouchers on educational inputs. *Unpublished working paper*, 2015.

C. E. Rouse. Private school vouchers and student achievement: An evaluation of the milwaukee parental choice program. *The Quarterly journal of economics*, 113(2):553–602, 1998.

C. E. Rouse and L. Barrow. School vouchers and student achievement: Recent evidence and remaining questions. *Annu. Rev. Econ.*, 1(1):17–42, 2009.

C. E. Rouse, J. Hannaway, D. Goldhaber, and D. Figlio. Feeling the florida heat? how low-performing schools respond to voucher and accountability pressure. *American Economic Journal: Economic Policy*, 5(2):251–81, 2013.

D. W. Schanzenbach. What have researchers learned from project star? *Brookings papers on education policy*, 9:205–228, 2006.

Y. Shi. Who benefits from selective education? evidence from elite boarding school admissions. *Economics of Education Review*, 74:101907, 2020.

V. Strauss. Three big problems with school 'choice'that supporters don't like to talk about. *Washington Post*, 2017.

B. Strøm and T. Falch. The role of teacher quality in education production. In *the Economics of Education*, pages 307–319. Elsevier, 2020.

C. Turner, E. Weddle, and P. Balonon-Rosen. The promise and peril of school vouchers. *NPR*, 2017. URL https://www.npr.org/sections/ed/2017/05/12/520111511/the-promise-and-peril-of-school-vouchers.

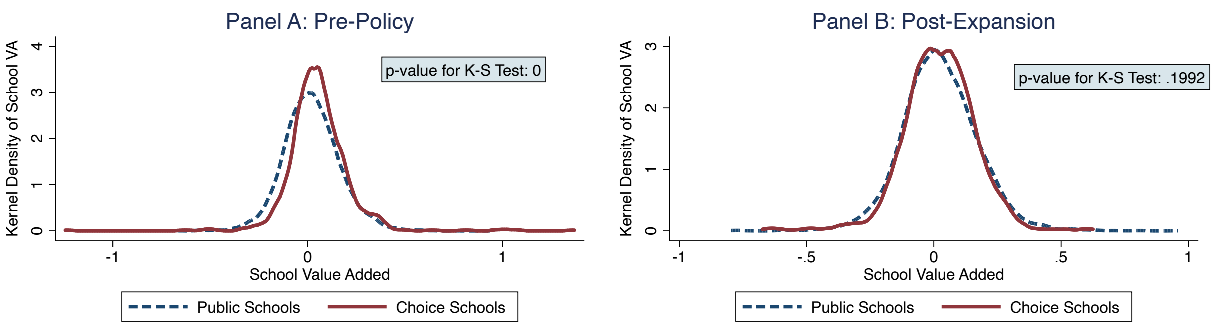
U.S Census Bureau. GS00ss08 - census bureau tables, 2021. URL `https://data.census.gov/cedsci/table?q=GS00SS08`.

R. J. Waddington and M. Berends. Impact of the indiana choice scholarship program: Achievement effects for students in upper elementary and middle school. *Journal of Policy Analysis and Management*, 37(4):783–808, 2018.

J. F. Witte, P. J. Wolf, J. M. Cowen, D. E. Carlson, and D. J. Fleming. High-stakes choice: Achievement and accountability in the nation's oldest urban voucher program. *Educational Evaluation and Policy Analysis*, 36(4):437–456, 2014.

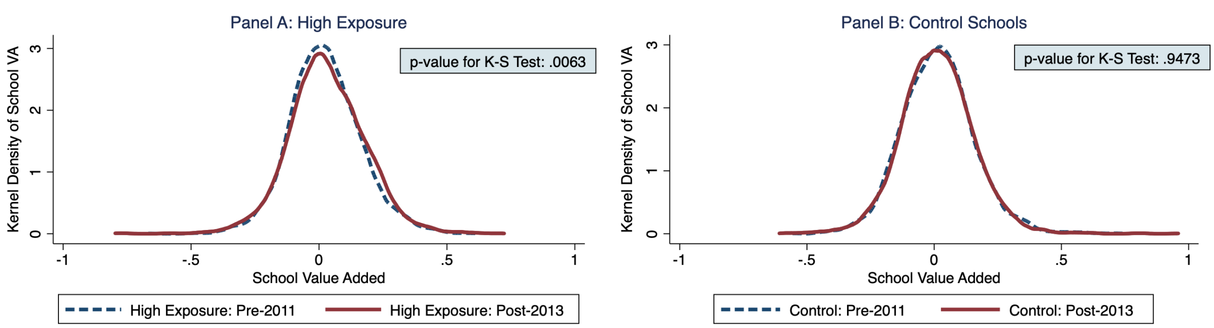
P. Wolf, B. Gutmann, M. Puma, B. Kisida, L. Rizzo, N. Eissa, and M. Carr. Evaluation of the dc opportunity scholarship program: Final report. ncee 2010-4018. *National Center for Education Evaluation and Regional Assistance*, 2010.

X Figures



Notes: This figure depicts the kernel density plots of our school value added (VA) estimates for the public and choice schools in our sample. Panel A shows the kernel density plots of schools in the years before the voucher program was implemented. Panel B shows those same estimates in the years after the program was expanded. School VA estimates are calculated using the OLS regression described by Equation (1). Data on test scores and enrollment come from the IDOE-CREO database. The p-value for the Kolmogorov-Smirnov equality-of-distributions test is reported.

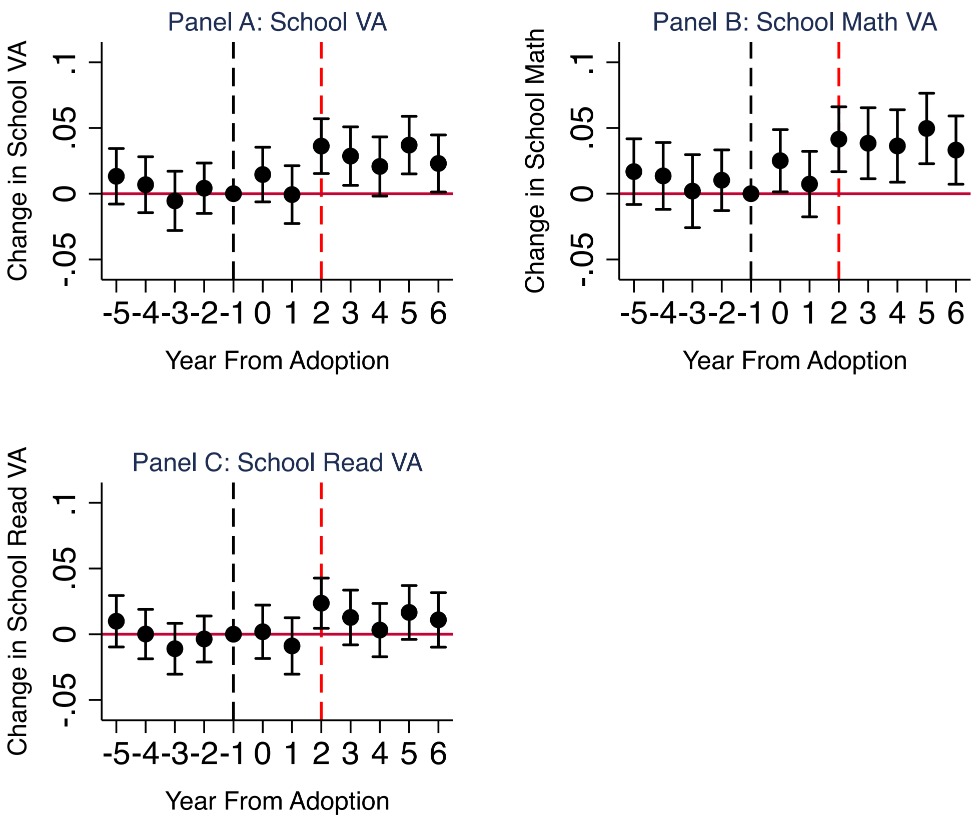
Figure 1.1: Kernel Density Plots of School VA - Public and Choice



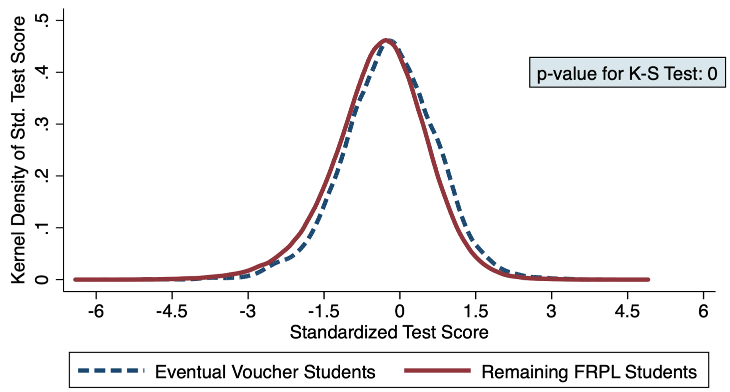
Notes: This figure depicts the kernel density plots of our school value added (VA) estimates for the public schools in our sample. Each panel plots school VA across two time periods: pre-2011 and post-2013 to align with the policy time horizons. Panel A shows the kernel density plots of schools facing high exposure to the policy. Panel B shows the kernel density plots for the control group. High exposure is defined as having a choice school within 5 miles of the school's location. School VA estimates are calculated using the OLS regression described by Equation (1). Data on test scores and enrollment come from the IDOE-CREO database. The p-value for the Kolmogorov-Smirnov equality-of-distributions test is reported.

Figure 1.2: Kernel Density Plots of School VA

Figure 1.3: Event-Study Results of Voucher Policy



This figure presents the event-study estimates from Equation (3). Figure 1.3(a) plots the estimates for overall school value-added, Figure 1.3(b) plots the estimates for school math value-added and Figure 1.3(c) plots the estimates for school reading value-added. Each figure is the result of a separate estimation. 95% confidence intervals are reported. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as having at least one nearby choice school. Data on enrollment come from the IDOE-CREO database and school value-added is calculated using Equation (1).



Notes: This figure depicts the kernel density plots of standardized test scores for the students attending public schools in the years before the voucher program. This figure shows the kernel density plots for the eventual voucher students and students remaining in the public school despite qualifying for a 90% voucher. The p-value for the Kolmogorov-Smirnov equality-of-distributions test is reported.

Figure 1.4: Kernel Density Plots of Standardized Test Scores

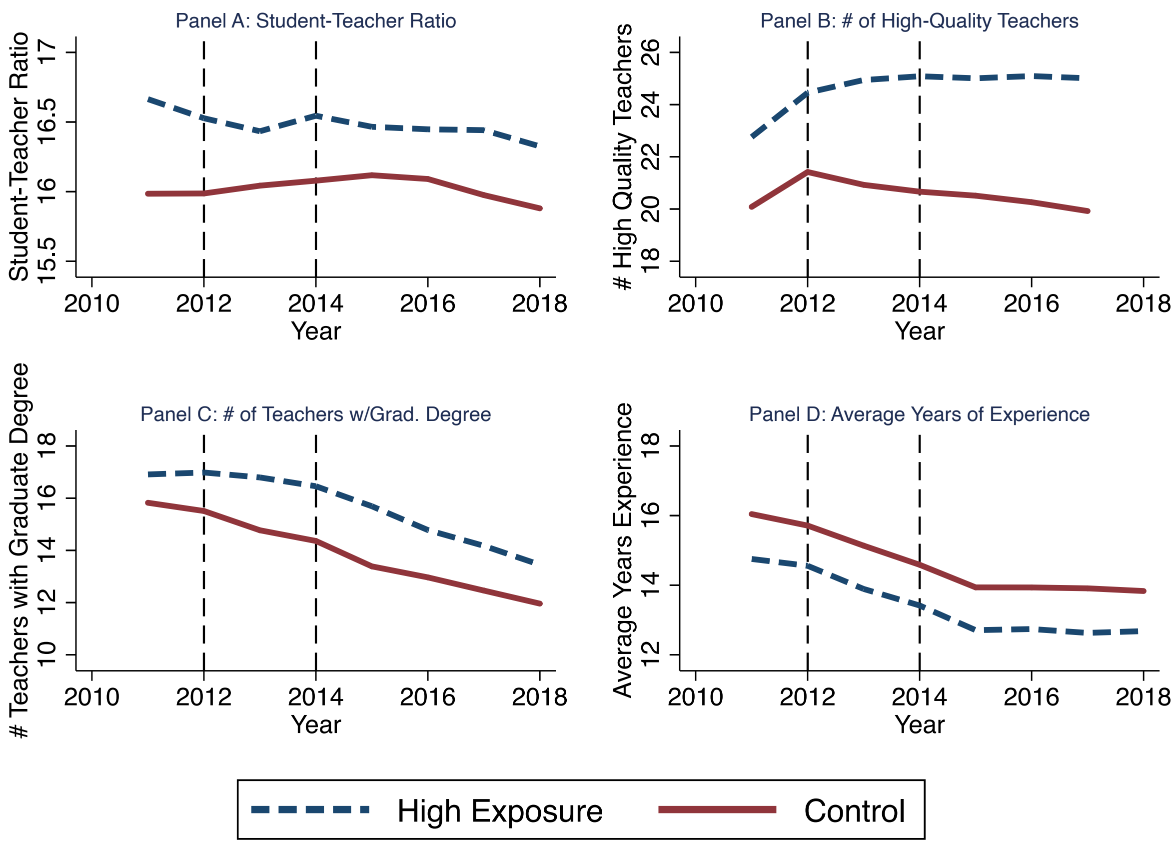
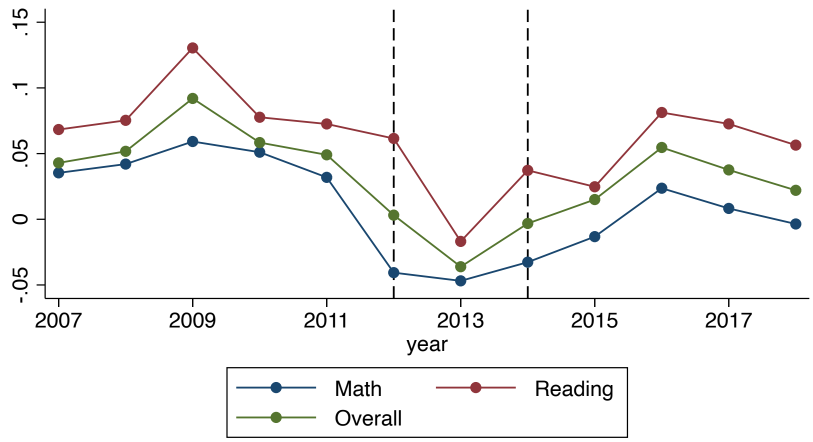


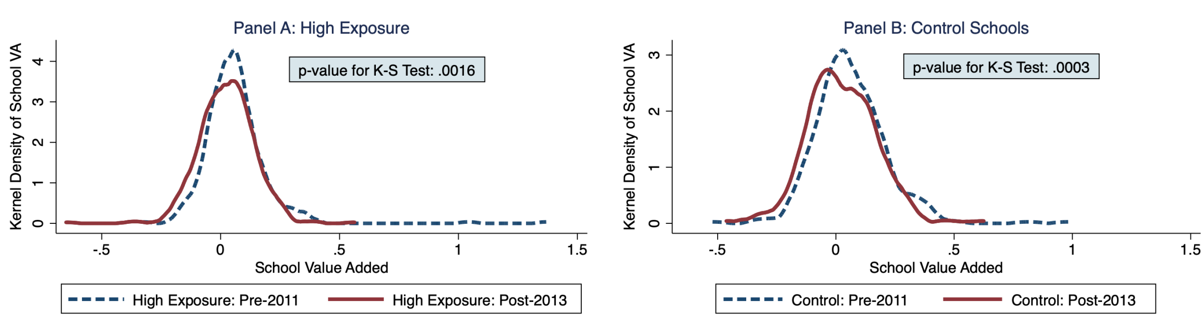
Figure 1.5: Public School Inputs

Notes: This figure presents the average student-teacher ratio (Panel A), number of high-quality teachers (Panel B), number of teachers with a graduate degree (Panel C) and average years of experience of teachers (Panel D) across public schools in the sample. High-exposure is defined as having a choice schools within five miles of the public school's location. Data on student-teacher ratios come from the Common Core of Data from the National Center of Education Statistics. Data on teacher characteristics come from the IDOE-Database.



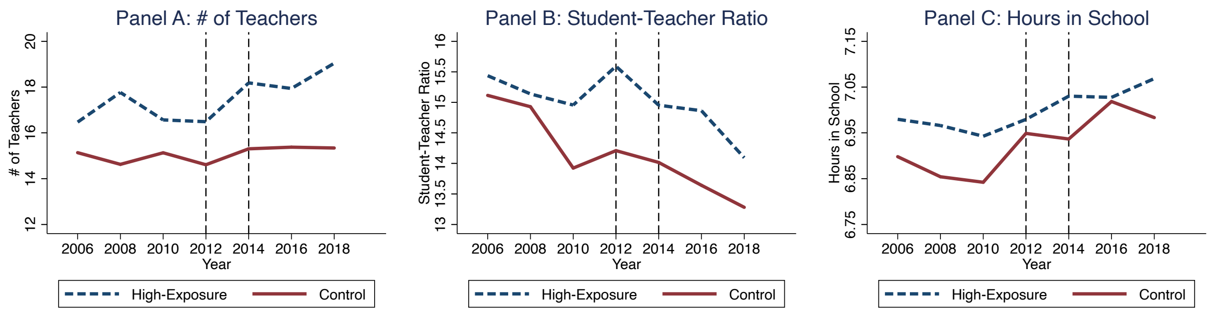
Notes: This figure depicts the average school value-added estimates across all choice schools in each year of the sample. School VA estimates are calculated using the OLS regression described by Equation (1). Data on test scores and enrollment come from the IDOE-CREO database. The dashed lines represents the years the voucher program was implemented and expanded.

Figure 1.6: Choice School Value-Added



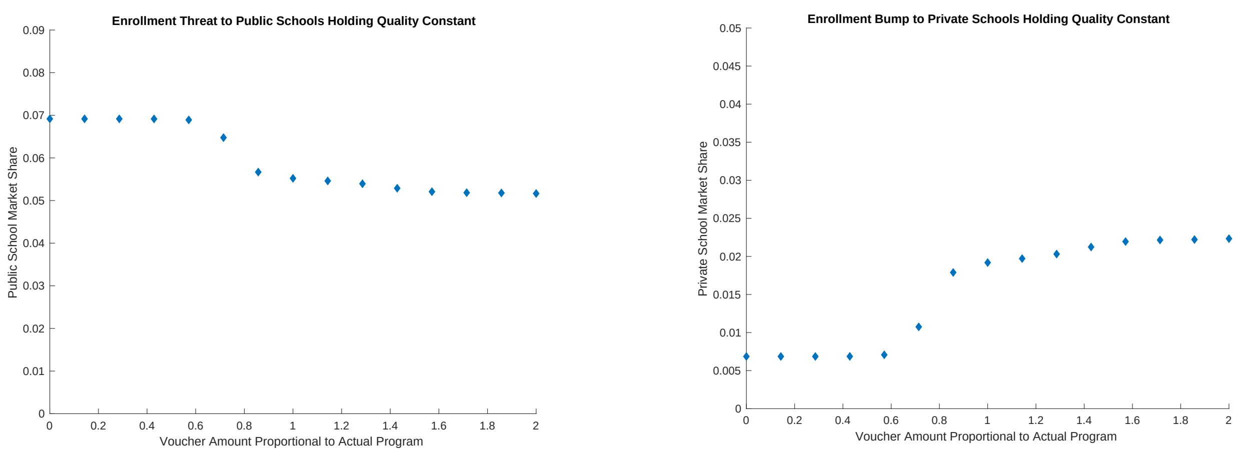
Notes: This figure depicts the kernel density plots of our school value added (VA) estimates for the public schools in our sample. Each panel plots school VA across two time periods: pre-2011 and post-2013 to align with the policy time horizons. Panel A shows the kernel density plots of schools facing high exposure to the policy. Panel B shows the kernel density plots for the control group. High exposure is defined as being in the top tercile of the distribution of number of public schools in 5-miles. School VA estimates are calculated using the OLS regression described by Equation (1). Data on test scores and enrollment come from the IDOE-CREO database. The p-value for the Kolmogorov-Smirnov equality-of-distributions test is reported.

Figure 1.7: Kernel Density Plots of School VA



Notes: This figure depicts the average number of teachers (Panel A), the average student-teacher ratio (Panel B), and the avereage hours spent in school (Panel C) across high-exposure and control choice schools in the sample. High-exposure choice schools are those in the top tercile of the distribution of number of public schools within 5 miles of the choice school's location. Data on choice school inputs come from the Private School Universe Survey conducted by the National Center for Education Statistics. Data are only avaiable in every other year. The dashed lines represent the years the voucher program was implemented and expanded.

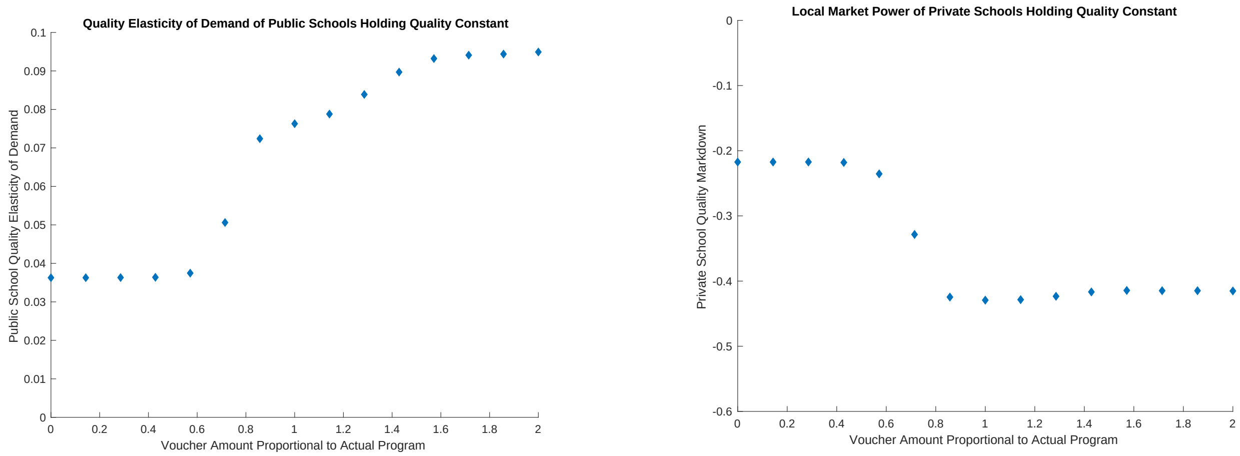
Figure 1.8: Choice School Inputs



(a) (b)

Figure 1.9: Changes in Enrollment Varying Voucher Amount

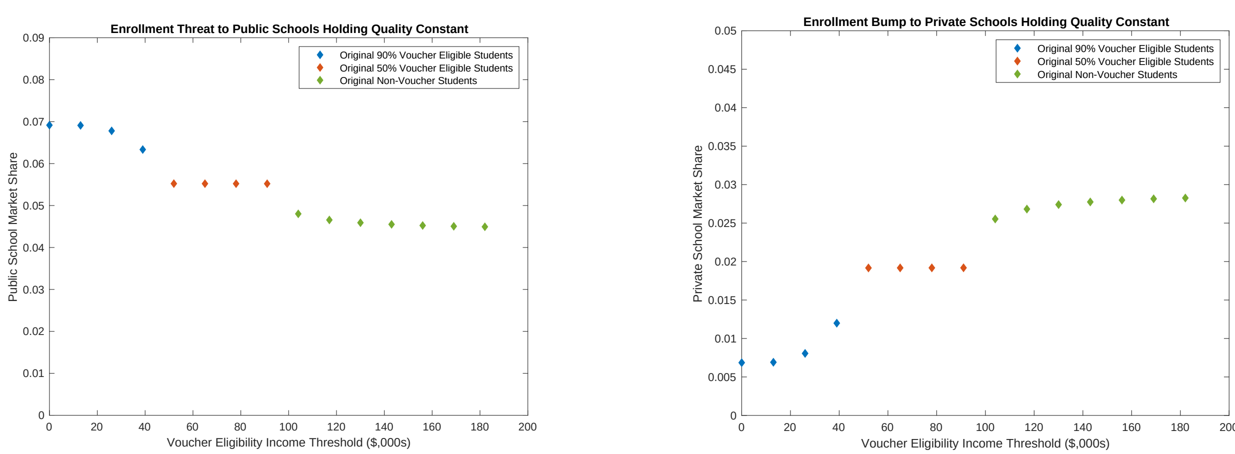
Notes: This figure presents the average market shares (enrollment) for public (Panel A) and private schools (Panel B) across our counterfactual voucher policy where we vary the voucher amount. The x-axis represents the multiplier we use to calculate the maximum voucher amount. A detailed description of our demand estimation used to created this figure can be found in Appendix Section A.5.



(a) (b)

Figure 1.10: Changes in Incentive to Provide Quality Varying Voucher Amount

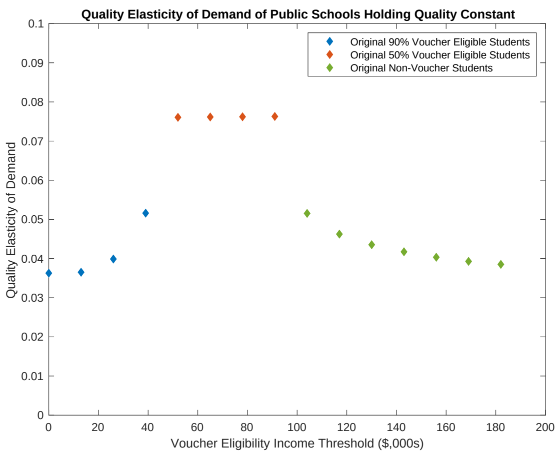
Notes: This figure presents the quality elasticity of demand for public schools (Panel A) and the quality markdowns for private schools (Panel B) across our counterfactual voucher policy where we vary the voucher amount. The x-axis represents the multiplier we use to calculate the maximum voucher amount. A detailed description of our demand and supply estimation used to created this figure can be found in Appendix Section A.5.



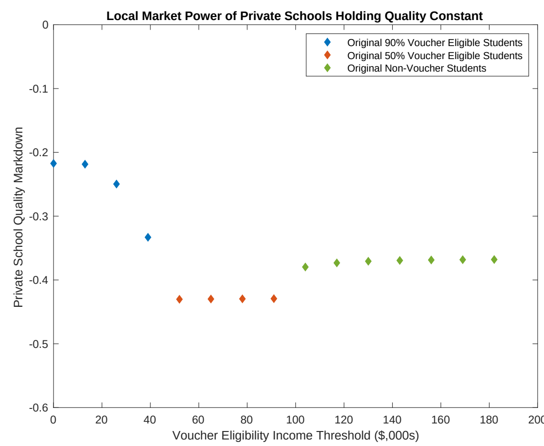
(a) (b)

Figure 1.11: Changes in Enrollment Varying Voucher Eligibility

Notes: This figure presents the average market shares (enrollment) for public (Panel A) and private schools (Panel B) across our counterfactual voucher policy where we vary the income threshold for voucher eligibility. The x-axis represents the maximum household income allowed to qualify for a voucher. A detailed description of our demand estimation used to created this figure can be found in Appendix Section A.5.



(a)



(b)

Figure 1.12: Changes in Incentive to Provide Quality Varying Voucher Eligibility

Notes: This figure presents the quality elasticity of demand for public schools (Panel A) and the quality markdowns for private schools (Panel B) across our counterfactual voucher policy where we vary the income threshold for voucher eligibility. The x-axis represents the maximum household income allowed to qualify for a voucher. A detailed description of our demand estimation used to created this figure can be found in Appendix Section A.5.

XI Tables

Table 1.1: Summary Statistics of High Exposure vs. Control Schools - Public

	(1) High Exposure	(2) Control	(3) Difference
# of Students Taking ISTEP+ Exam	262	217	45***
	(241)	(176)	
School VA	0.021	0.018	.003
	(0.146)	(0.153)	
School Math VA	0.025	0.026	.001
	(0.150)	(0.143)	
School Reading VA	0.009	-0.002	.011
	(0.135)	(0.133)	
% White	0.648	0.914	-0.265***
	(0.271)	(0.121)	
% Black	0.156	0.017	0.139***
	(0.200)	(0.097)	
% FRPL	0.550	0.415	0.135***
	(0.257)	(0.155)	
N	727	552	

This table presents summary statistics for the set of schools identified as high exposure and the control group in the year before the voucher policy was implemented. High exposure is defined as having at least one nearby choice school. Column (3) denotes the difference in the means between schools in the control group and those highly exposed to the program. Data on enrollment come from the IDOE-CREO database and school value-added is calculated using Equation (1).

Table 1.2: Summary Statistics of High Exposure vs. Control Schools - Choice

	(1) High Exposure	(2) Control	(3) Difference
# of Students Taking ISTEP+ Exam	145	107	38**
	(89)	(73)	
School VA	0.056	0.059	.003
	(0.197)	(0.123)	
School Math VA	0.048	0.052	.004
	(0.254)	(0.170)	
School Reading VA	0.082	0.076	-.006
	(0.153)	(0.105)	
% White	0.749	0.904	0.154***
	(0.271)	(0.102)	
% Black	0.077	0.013	-0.065***
	(0.158)	(0.389)	
% FRPL	0.264	0.101	-0.163***
	(0.287)	(0.103)	
N	54	124	

This table presents summary statistics for the set of schools identified as high exposure and the control group in the year before the voucher policy was implemented. High exposure is defined as having at least one nearby choice school. Column (3) denotes the difference in the means between schools in the control group and those highly exposed to the program. Data on enrollment come from the IDOE-CREO database and school value-added is calculated using Equation (1).

Table 1.3: DiD Results on the Effects of High Exposure on School VA

	(1) School Value-Added	(2) School Value-Added	(3) School Math Value-Added	(4) School Math Value-Added	(5) School Reading Value-Added	(6) School Reading Value-Added
$Post_t \cdot HighExp_s$	0.023***	0.009	0.030***	0.015*	0.013***	-0.000
	(0.006)	(0.006)	(0.007)	(0.008)	(0.005)	(0.006)
Interaction with High Share of FRPL in 2010		0.030***		0.032***		0.028***
		(0.008)		(0.010)		(0.007)
Observations	15,348	15,348	15,348	15,348	15,348	15,348
R-squared	0.448	0.449	0.433	0.434	0.455	0.456
Baseline Mean	0.0197	0.0197	0.0255	0.0255	0.00390	0.00390

Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$. All standard errors are clustered at the school level. Each coefficient is the result of a separate estimation. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as having at least one nearby choice school. Columns (2), (4), and (6) include the interaction of high exposure and an above median share of FRPL students in the year before the voucher policy was implemented. Data on enrollment come from the IDOE-CREO database and school value-added is calculated using Equation (1).

Table 1.4: DiD Results on Public School Inputs

	(1) Student-Teacher Ratio	(2) # of Teachers w/Grad. Degree	(3) # of HQ Certified Teachers	(4) Avg. Years of Experience
$Post_t \cdot HighExp_s$	0.11	0.68***	1.48**	-0.090
	(0.15)	(0.24)	(0.31)	(0.15)
Observations	9,963	10,179	10,179	10,179
Baseline Mean	18.34	16.91	22.88	14.75

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All standard errors are clustered at the school level. Each coefficient is the result of a separate estimation. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as having at least one nearby choice school. Observations are lower compared to other tables because of limited availability of data. Data on teacher characteristics come from the IDOE-CREO database and student-teacher ratios are calculated from the Common Core of Data on Public Schools.

Table 1.5: DiD Results on Attendance and Suspension Measures

VARIABLES	(1) Percent Days Attend	(2) Total Days Attend	(3) Percent Expelled or Suspended
$Post_t \cdot HighExp_s$	0.327***	0.517**	-0.056
	(0.116)	(0.215)	(0.156)
Observations	15,336	15,336	15,336
R-squared	0.866	0.860	0.769
Baseline Mean	88.62	159.5	5.594

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All standard errors are clustered at the school level. Each coefficient is the result of a separate estimation. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as having at least one nearby choice school. Data on enrollment come from the IDOE-CREO database. One school is missing data for all years on attendance, so the number of observations is slightly less than other tables.

Table 1.6: DiD Results on Demographics of Students Enrolled

VARIABLES	(1) Share Female	(2) Share White	(3) Share Black	(4) Share Hispanic	(5) Share FRPL
$Post_t \cdot HighExp_s$	-0.188	-2.719***	0.267*	2.268***	0.383
	(0.211)	(0.250)	(0.151)	(0.204)	(0.354)
Observations	15,348	15,348	15,348	15,348	15,348
Baseline Mean	49.58	76.30	9.607	8.082	49.20

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All standard errors are clustered at the school level. Each coefficient is the result of a separate estimation. All regressions include school and year fixed effects. Baseline covariates include the share of students that are section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as being in the top tercile of the distribution of number of nearby public schools. Data on enrollment come from the IDOE-CREO database.

Table 1.7: DiD Results on Predicted School Value-Added

VARIABLES	(1) Predicted School VA	(2) Predicted School Math VA	(3) Predicted School Reading VA
$Post_t \cdot HighExp_s$	-0.010**	0.003	-0.023***
	(0.004)	(0.006)	(0.003)
Observations	15,348	15,348	15,348
R-squared	0.354	0.274	0.427

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the school level. Each coefficient is the result of a separate estimation. Regressions include school and year fixed effects. High exposure is defined as having at least one nearby choice school. Data on test scores come from the IDOE-CREO database. Predicted School Value-Added are estimated by regressing value-added in 2007 on school characteristics and using the regression coefficients to predict school-value added for all years in the sample.

Table 1.8: Heterogenous DiD Results of Voucher Program

VARIABLES	(1) School Value-Added	(2) School Math Value-Added	(3) School Reading Value-Added
Panel A: Large Baseline Enrollment			
$Post_t \cdot HighExp_s$	0.029***	0.036***	0.019***
	(0.008)	(0.010)	(0.007)
Interaction with Above Median Baseline Enrollment	-0.011	-0.013	-0.010
	(0.007)	(0.009)	(0.006)
Observations	15,348	15,348	15,348
R-squared	0.448	0.434	0.455
Baseline Mean	0.0197	0.0255	0.00390
Panel B: High Baseline School Value-Added			
$Post_t \cdot HighExp_s$	0.045***	0.059***	0.029***
	(0.007)	(0.008)	(0.006)
Interaction with Above Median Baseline School VA	-0.041***	-0.053***	-0.030***
	(0.007)	(0.009)	(0.006)
Observations	15,348	15,348	15,348
R-squared	0.450	0.436	0.456
Baseline Mean	0.0197	0.0255	0.00390
Panel C: Above Median Neighborhood Income			
$Post_t \cdot HighExp_s$	0.027***	0.032***	0.019***
	(0.007)	(0.009)	(0.006)
Interaction with Above Median Neighborhood Income	-0.007	-0.005	-0.011
	(0.008)	(0.009)	(0.007)
Observations	15,348	15,348	15,348
R-squared	0.448	0.433	0.455
Baseline Mean	0.0197	0.0255	0.00390

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All standard errors are clustered at the school level. Each coefficient is the result of a separate estimation. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as having at least one nearby choice school. A small (big) school is defined as one that falls below (above) the median in total enrollment in the 2006-2007 AY. A low (high) baseline VA school is defined as one that falls below (above) the median in VA in the 2006-2007 AY. Data on enrollment come from the IDOE-CREO database and school value-added is calculated using Equation (1).

Table 1.9: DiD Results Using Choice Schools

VARIABLES	(1) School Value-Added	(2) School Math Value-Added	(3) School Reading Value-Added
$Post_t \cdot HighExp_s$	-0.010	-0.012	-0.014
	(0.018)	(0.022)	(0.014)
Observations	2,136	2,136	2,136
R-squared	0.295	0.312	0.300
Baseline Mean	0.0490	0.0319	0.0725

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All standard errors are clustered at the school level. Each coefficient is the result of a separate estimation. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as being in the top tercile of the distribution of the number of nearby public schools. Data on enrollment come from the IDOE-CREO database and school value-added is calculated using Equation (1).

Table 1.10: DiD Results on Choice School Inputs

	(1) Full-Time Teachers	(2) Student/Teacher Ratio	(3) Hours in School Day
$Post_t \cdot HighExp_s$	-0.342	0.930**	-0.080
	(0.893)	(0.418)	(0.051)
Observations	977	977	977
Baseline Mean	15.57	14.24	6.873

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All standard errors are clustered at the school level. Each coefficient is the result of a separate estimation. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as being in the top tercile of the distribution of the number of public schools within five miles. Data on choice school inputs comes from the Private School Universe Survey which is conducted biannually. There are fewer observations in this analysis because of the survey design.

2. THE EFFECT OF FIELD TRAINING OFFICERS ON POLICE USE OF FORCE

Over the past decade, police use of force has become an increasingly charged political issue with growing calls for reform. One of the few reforms where advocates and the policing community have reached a consensus is on the need for improved and expanded training. In this paper, we study an under-researched but nearly universal training approach whereby a recruit is paired with a senior officer during a phase referred to as "field training". In particular, we consider the link between a field training officer's prior propensity to use force and a recruit's subsequent enforcement behavior. We leverage a unique setting where recruits are as-good-as-randomly assigned to field training officers and where we have detailed information on the universe of calls for service. We document meaningful differences across field training officers in terms of their propensity to use force prior to being paired with a recruit. Further, we find that a one standard deviation increase in a field training officer's propensity to use force (138 percent) is associated with a 12 percent increase in their recruit's subsequent propensity to use force. The effect of having a more aggressive field training officer persists for as much as two and a half years after the recruit completes training.

Correspondence: Adger, Texas A&M University, cadger@tamu.edu; Ross, Northeastern University, ma.ross@northeastern.edu; Sloan, United States Military at West Point, carlywill.sloan@westpoint.edu

Acknowledgments: We are grateful for useful comments from Jeffery Grogger, Elizabeth Luh (twice!), Aurelie Ouss, Steve Ross, Jamein Cunningham, the Dallas Police Department, Steve Mello, John MacDonald, Shawn Bushway, Anne Piehl, and participants at the 2022 NBER Summer Institute, 2022 Al Capone Conference, 2022 Chicago/LSE Conference, 2021 Southern Economic Association Conference as well as the Association for Public Policy & Management Conference. Sloan and Ross are also grateful to the Russell Sage Foundation (G-2011-30016) for financial support of our project "Does More Training Mitigate Disparities in Police Use of Force?". We are also grateful to Felipe Goncalves and Emily Weisburst for sharing DPD overtime and shift data with us.

I Introduction

Over the past decade, police use of force has become an increasingly charged political issue with growing calls for reform. As of 2020, 65 percent of Americans believe that police officers are using an inappropriate amount of force (Pew Research Center, 2020). Concerns about appropriate use of force, combined with high-profile killings of unarmed Black individuals by police, has led most Americans to support the need for reform (Gallup Panel, 2020). However, there is substantial disagreement amongst policymakers and the public about how best to implement policing reform. For example, the well-known "defund the police" movement is supported by only 31 percent of Americans (Rakich, 2020) while policies like eliminating the enforcement of nonviolent crimes are only supported by 50 percent (Gallup Panel, 2020). However, polls have found that nearly 90 percent of Americans support improved and expanded training in areas like de-escalation and avoiding violence (Ipsos, 2021). Policymakers and the policing community also frequently cite training as a key approach to reducing police violence. In fact, the need for more and better police training is one of the few areas where the public, advocates, and the policing community can agree on a potential policy solution.

The small literature that has sought to evaluate the effectiveness of police training has historically relied on surveys of officers about the impact of the training following participation in a particular course. Despite the perceived importance of training, there have been few empirical studies that consider the impact of police training on actual enforcement outcomes, such as use of force. Further, the few studies that have examined the impact of training on actual enforcement outcomes have exclusively focused on classroom-based training interventions occurring at the police Academy, or as a part of continuing education requirements (McLean et al., 2020, Johnson et al., 2021, Owens et al., 2018). Thus, there is little work on the effectiveness of experiential components of training, broadly referred to as "field training", on police behavior and enforcement outcomes. Apprenticeship-style models of field training are nearly universal among policing agencies in the United States and consist of an inexperienced recruit, who has graduated the police academy, partnering with a more experienced officer on patrol for about six months. Although the few studies that have focused on classroom style training have found little to no long-term effects on subsequent enforcement behavior, there are two key reasons to believe that field training might have a larger and more persistent impact. First, the law enforcement community generally characterizes their field training as "the most important stage in the process of becoming an independent police officer" (Warners, Ronald, 2020). Second, recent work by Holz et al. [2020), West [2019), Weisburst [2022) suggests that an officer's on-the-job-experiences are a significant factor in their subsequent enforcement behavior.

In this paper, we ask whether recruits who are assigned to more aggressive field training officers (FTOs) are subsequently more likely to use force in the years following field training. To address this question, we leverage administrative data on calls for service (i.e. 911 calls) from the Dallas Police Department covering a period from 2013 to 2019. In this particular institutional setting, there is as-good-as-random assignment of recruits to training officers over the sample period. We believe that this quasi-random variation closely mirrors the ideal experiment, and allows us to identify the causal effect of being assigned to an FTO, who has previously been more likely to use force, on a recruit's subsequent enforcement behavior.We characterize aggressive FTOs by constructing a measure of propensity to use force in the period prior to their being assigned a particular recruit. In constructing FTO force, we account for the fact that some FTOs may be assigned or select more dangerous calls or locations by considering factors such as neighborhood, characteristics of the call, and aspects related to date and time. Thus, this measure allows us to assess whether a recruit is assigned to a particular FTO who has historically responded to a similar set of incidents as their peers but has behaved more aggressively.

In our preferred specification, we estimate that a one standard deviation increase (138 percent) in an FTO's prior propensity to use of force is associated with a 12 percent increase in a recruit's subsequent likelihood of using force after training. Distinct from other types of training model that have been studied previously in the literature, we find that the effect of having a more forceful FTO is particularly long-lasting. Namely, we show that effects persist for at least two and a half years after the completion of training. We interpret our results as being consistent with the idea that having a more forceful FTO, results in a recruit who subsequently has a lower threshold for using force on any given call for service. We are also able to rule out several alternative mechanisms. In particular, we document that recruits assigned to a more forceful FTO are not involved in more active forms of policing as captured through arrest rates, response times, and time spent on a given call. Further, we document that our findings are not driven by (1) recruits witnessing a force incident during field training, (2) the propensity of a recruit to subsequently answer a call for service with their FTO or another partner after training, (3) differences in report-writing behavior, and (4) other FTO characteristics which are correlated with force. Last we show that recruits assigned a higher force FTOs are more likely to issue less serious or unfiled arrests, but there is no change in felony or filed arrests. We argue that these results are generally consistent with the story that less desirable use of force behavior is passed from high force FTOs to recruits.

Our main finding, that a recruit's subsequent use of force is shaped by their FTO, contributes to a broad literature that spans criminology to economics and emphasizes the characteristics of individual police officers in shaping police use of force. Criminologists were among the first to investigate heterogeneity within policing agencies (Crank, 2004, Chan, 1997, Paoline III, 2006, Bruinsma and Weisburd, 2014, Woody, 2005). Recent work in economics has focused on the role that an officer's race or their peer's race has on their propensity to use force (Ba et al., 2021, Weisburst, 2019, Fryer Jr, 2019, Hoekstra and Sloan, 2020, Rivera, 2022). Our findings also contribute to a broader literature in criminology about the impact of policing culture. Drawing on an established literature (e.g., Skolnick, 1966, Westley, 1970, Paoline III, 2003), the President's Task Force on 21st Century Policing [2015) recommends that policing agencies transition from a "warrior" to "guardian" culture and states that "Field Training Officers impart the organizational culture to the newest members." Paoline III [2003) identifies the field as one of the most important settings for the transfer of policing culture, and Paoline III and Gau [2018) state that culture is key to reducing aggressive policing. From a policy perspective, our findings represent the first rigorous empirical evidence supporting interventions aimed at reducing force by changing police culture particularly through field training. Further, our study suggests that characteristics like aggression or prior enforcement behavior, which are typically not documented or explicitly tracked by policing agencies, are key sources of officer heterogeneity and important determinants of aggregate rates of force.

We also add to the literature on the influence of social interactions in the workplace. This literature documents that, across many different settings, social influence by peers and superiors shape how an individual agent makes decisions. Many studies have demonstrated that peers are an important determinant of future decision making across a variety of settings such as school (e.g., Sacerdote, 2001, Whitmore, 2005, Carrell et al., 2013, Anelli and Peri, 2019, Bifulco et al., 2011), work (Mas and Moretti, 2009), neighborhoods (Glaeser et al., 1996, Billings et al., 2019, Billings and Schnepel, 2020), and the military (Murphy, 2019). A distinct but related literature also documents the importance of teachers and managers in altering the subsequent behavior of students and employees (see Bertrand and Schoar, 2003, Bloom and Van Reenen, 2007, Lazear et al., 2015, Giorcelli, 2019, Fenizia, 2021). Our work contributes to both of these fields of research by considering the role of a particularly influential peer/superior in an extremely high-stakes setting, i.e. policing and use of force. Our finding also emphasizes the importance of policy interventions that alter the composition of one's peers or their supervisor, particularly during intensive periods of on-the-job training.

The two prior works closest to our own are Holz et al., 2020 and Getty et al., 2014. In Holz et al. [2020), the authors examine the impact of an officer's peer being injured on the job and find that other officers are more likely to use force, receive complaints, and injure suspects in the week following their peer's injury. We differ from Holz et al., 2020 by considering a fundamentally different determinant of force (i.e. FTO force propensity). Relative to a peer getting injured on the job, we find that one's FTO has a substantially longer impact on subsequent policing behavior. Namely, we document higher use of force for officers trained by more aggressive officers for as much as three years after the completion of their training. In Getty et al., 2014, the authors also use data from the Dallas Police Department to study the impact of FTOs on subsequent allegations of misconduct. Using hierarchical modeling, the authors find that 26.5% of the overall variation in a recruit's subsequent complaints are related to their FTO. They do not ask whether having an FTO with more complaints causes a recruit to acquire more complaints. Relative to Getty et al., 2014, we implement a more rigorous empirical design which allows us to obtain plausibly causal estimates to address an entirely different conceptual question, i.e. the relationship between an FTO's and a recruit's propensity to use force.

The findings of this paper have several important policy implications. First, by demonstrating that FTOs are an important determinant to a recruit's subsequent propensity to use force, we have identified a particularly promising target for policy interventions aimed at reducing aggregate rates of police use of force. To our knowledge, there has never been any targeted interventions aimed at substituting high for low-force FTOs in an effort to reduce aggregate rates of police force. Given our findings and the fact that classroom based procedural justice interventions (McLean et al., 2020, Owens et al., 2018, Wheller et al., 2013) have lead to large but only short-lived effects on enforcement behavior, there is reason to believe that reforms to field training might have larger and longer-lasting impacts on aggregate force. [1] Second, our findings are conservative in that we do not attempt to explore the dynamic effects of changing the composition of FTOs. In particular, our findings suggest that selecting less-forceful FTOs in any given year would lead to less forceful recruits who will then subsequently become FTOs in the future. Thus, the dynamic effects of altering the composition of FTOs in any given year will potentially have large effects on aggregate force that will last very far into the future. Third, changing the composition of FTOs is a much more cost-effective (even potentially costless) intervention relative to alternatives that involve classroom or simulation based training. Finally, reforms aimed at changing the composition of FTOs are much more practical and politically palatable relative to large-scale initiatives such as defunding the police, dramatically reducing enforcement policies, or attempts to substantially alter

[1] Procedural justice refers to the idea of fairness in the processes that resolve disputes and allocate resources. It is a concept that, when embraced, promotes positive organizational change and bolsters better relationships. Procedural justice speaks to four principles, often referred to as the four pillars: fairness in the processes, transparency in action, opportunities for voice, impartiality in decision making (COPS, 2020). There is also a body of literature that focuses on the impact of short procedural justice training sessions on officer attitudes (e.g., McLean et al., 2020, Rosenbaum and Lawrence, 2017, Schaefer and Hughes, 2016, Skogan et al., 2015. We differ from this literature because of our focus on officer actions.

the composition of the police force.

II Police Officer Training and Institutional Background

According to a 2018 survey of 681 state and local law enforcement agencies, police recruits spent an average of 833 hours in basic training (at the Academy) and 508 hours in field training (Bureau of Justice Statistics, 2018). Thus, we believe that the findings from our study are broadly generalizable because the training received by recruits in Dallas is representative of how the vast majority of policing agencies across the country train new officers. However, we note that recruits in Dallas tend to receive more training (both at the Academy and in the field) than the average agency, and that the requirements exceed those mandated by the State of Texas. Like most agencies, however, training in Dallas is divided into two distinct phases before recruits become a full-fledged police officer, i.e. Academy training (phase 1) and field training (phase 2).[2] Although our paper focuses exclusively on the field training component of a recruit's preparation for becoming a police officer, we provide a brief but comprehensive discussion of the full training process in this section.

In the first phase of training, recruits must graduate from the Dallas Police Academy. Training at the Academy lasts at least 36 weeks and consists of 1,431 hours of instruction. At the beginning of their time at the Academy, recruits undergo mental and physical training aimed at preparing them for the demands of a career in law enforcement. Next, the recruits complete legislatively mandated classroom and scenario-based training as well as a number of additional courses required by Dallas PD. The legislatively mandated courses are developed by the Texas Commission on Law Enforcement (TCOLE) which is the regulatory agency governing the licensure of all peace officers in Texas. TCOLE also regulates subsequent in-service training requirements which are necessary to maintain a peace officer license in Texas. In most states, there is a similar governing agency (known colloquially as "Post", i.e. Police Officer Standards & Training) which sets both Academy and in-service training requirements. Although there is some variation in the specific training required in different states, a national organization (the International Association of Di-

[2] Police Officer is the lowest rank in the Dallas Police Department

rectors of Law Enforcement Standards and Training) issues a core set of recommendations which have been broadly implemented across the country and are consistent with how Dallas PD trains police officers.

After completing basic training at the Academy, recruits enter a second phase of training referred to as "field training". As noted above, this second phase of training is the focus of our study and has largely been overlooked by the existing empirical literature on policing. During field training, recruits ride with more a experienced officer (i.e. their FTO) in, what could be characterized as, an apprenticeship style model where they are gradually afforded more autonomy. FTOs have a dual responsibility of providing service in their sector while simultaneously providing on-the-job training for their assigned recruits. At the end of field training, recruits are evaluated by their FTO and, based on a successful evaluation, graduate to becoming full-fledged police officers. In our setting, nearly all recruits who successfully graduate from the Academy also successfully complete field training and virtually all of those officers are initially assigned to patrol, i.e. answering calls for service. This apprenticeship style model of on-the-job training was first developed by the San Jose Police Department in the 1960s and has since become a near universal standard for how law enforcement agencies in the United States approach the training of new recruits.

Key to the empirical design of our study, recruits in Dallas have no discretion in choosing their FTO. In particular, recruits are assigned to one of seven divisions in Dallas to complete their field training. This assignment is based on the staffing needs of the division rather than the skills or performance of recruits at the Academy. Within a given division, recruits are then assigned to FTOs and their associated patrol sectors/beats. Here, command staff at Dallas PD indicated that these decisions are entirely unrelated to recruit characteristics or their performance at the Academy. In a subsequent section, we provide empirical evidence supporting the claim that, conditional on division, the initial assignments provide as-good-as random variation in the pairing of recruits to FTOs.

In Dallas, the field training process takes a total of six months to complete and consists of four phases. In the first and fourth phase, a recruit is paired with the same FTO. In the second and third phase, the FTO is different. The first three training phases of field training are seven weeks long. The final evaluation phase is conducted by the initial (i.e. phase one) FTO and lasts three weeks. When field training begins, recruits are instructed to take on a more observational role. As training progresses, they are given more autonomy and become an active participant in responding to calls for service. For example, in the early weeks of field training, a recruit may simply watch a FTO respond to a call for service. In later phases, the recruit may lead the response under the guidance and observation of their FTO. FTOs also conduct frequent, often daily, evaluations of recruits. According to command staff in Dallas, these evaluations are largely used to provide the recruit with extensive feedback on their performance.

After field training is complete, recruits then spend another year on probation where they are required to stay in their initial division assignment and associated sector. During the first six months of probation, commonly called "little t" by Dallas command staff, recruits are required to choose a more experienced officer to ride with as their partner. Finally, one year after completing the Academy, recruits are taken off probation and advance to the position of Police Officer.

This paper focuses on the impact that the first FTO has on the recruit's subsequent enforcement behavior. We made the decision to focus on the first FTO for two reasons. First, in our conversations with Dallas police officers, they communicated that field training shapes officers' policing “style" much more than their training at the Academy. Command staff in Dallas also emphasized that this phase of training is the most critical part of a recruit's development and that all peace officers remember the lessons learned during field training for the rest of their career. Second, police officers and command staff in Dallas noted that the first phase of field training is the most significant because it is a recruit's first exposure to providing service. Further, recruits often return to their initial FTO for their final training and evaluation phase.[3] We also focus on FTOs rather than the officer a recruit chooses to work with during "little t" because FTOs are conditionally randomly assigned. In contrast, a recruit may select their partner during "little t."

[3] If we estimate the effect of each FTO in the same regression, only the force rate of a recruit's first FTO has a large and statistically significant effect (see Table B.1). We interpret this as evidence consistent with our conversations with Dallas officers.

In our study, we document FTO and recruit behavior using 911 calls for service. When a civilian calls 911 in Dallas, they are first connected to a 911 operator. The operator will then record essential characteristics of the call such as location, description of events, and time in the Computer Aided Dispatch System (CAD). The operator will also place the call into a standardized category, such as "domestic disturbance." Finally, the operator also records their perception of the urgency and severity of a call. This is referred to as the priority of the call and is assigned values from 1 to 4, with 1 being the highest priority. The information recorded in CAD system is then provided to police dispatchers whose job it is to assign calls to police officers. Dispatchers assign calls to officers based on priority level (relative to other calls in the queue), proximity, and availability. If there are many more active calls than available officers, lower-priority calls are postponed until higher priority calls are resolved. Dispatchers also decide the number of officers to assign to a call based on Dallas PD's standards about how many officers are required for different types of calls. For example, more serious incidents (such as shootings and mental health calls) may involve the dispatch of multiple officers. Once an officer responds to a call, the officer is afforded a significant amount of discretion in how they handle an incident in terms of their decision to make an arrest or use force.

To measure officers use of force, we link 911 calls to force reports.[4] In general, the way that Dallas PD tracks force incidents is consistent with best practices established by criminologists and embraced by many law enforcement agencies across the county. In particular, Dallas PD officers are required to make a "Response to Resistance" entry in a proprietary database called BlueTeams. following a force incident[5] All force incidents are reviewed by a supervisor (Dallas Police Department General Orders, 2015). According to the Dallas Police "The physical control techniques used may range from the use of handcuffs in an arrest, strikes with an impact weapon, or

[4]Dallas refers to force as a response to resistance.

[5]Any Response to Resistance that is Soft Empty Hand Control or above on the Response Continuum, with the exception of "Compliant Handcuffing" only. This will include, but not be limited to the following: 1. All take-downs, pressure points, joint locks 2. Any use of Oleoresin Capsicum Chemical Spray. 3. Any deployment of the Pepperball System. 4. Personal weapons such as hands and feet. 5. Any use of the baton or any other type of instrument that is used as an impact weapon. 6. Any use of an Electronic Control Weapon (Taser). This includes accidental discharges of the Taser. 7. The deployment of a firearm which is pointed directly at any individual.

the use of a firearm" (Dallas Police Department, 2019). According to police officers and command staff at Dallas PD, the penalty for not correctly reporting a force incident is extremely severe and compliance is virtually universal. We also link 911 calls to arrest reports. Here we observe the type of arrest made (felony, misdemeanor, or n-class) and well as demographics of the arrested (race, gender, age). We also categorize felony and misdemeanor arrests as filed or unfiled. If an arrest is unfiled then the district attorney decided to not move forward with the case and the defendant will not be charged with a crime. We interpret unfiled arrests as resulting from lower quality police work. [6]

III Data and Summary Statistics

III.A Analytical Sample

Our analytical sample is derived from the universe of 3.9 million calls for service (i.e. 911 calls) received by Dallas PD from Jan 2013 to July 2019. We link these data to force reports, arrest records, Dallas County District Attorney records, and officer characteristics.[7][8] According to the Dallas Police Department, they do not keep an official historical list of recruit-FTO pairings for each of the four field training phases. However, we have been provided detailed information on the dates of specific assignments for each officer in our sample as well as Academy graduation dates. Thus, we are able to construct recruit-FTO pairings for each field training phase using these dates as well as the likelihood a recruit arrives to a call with a senior police officer.[9] In particular, we construct a set of dates for each recruit which are associated with each phase of field training.

[6] Most "n-class" arrests are made for warrants.

[7] In linking the force records with calls for service, we do so based on the incident identifier but not the officer badge number. Although our results are generally robust to linking this data on both incident and badge number, we have taken a conceptual stance that it is more correct to associate an incident resulting in force with every officer on the scene. This is because one officer may influence another officer on the scene. We also restrict force incidents to those we are confident (based on the time stamp) occurred on the scene of an incident as opposed to those occurring after a suspect is in custody. We attach arrests to calls for service in a similar fashion such that all officers on the scene are associated with a given arrest regardless of whether they are the specific officer listed in the arrest report.

[8] We link arrests to Dallas District Attorney Data on filed cases using defendant name and date of offense.For each match, we block on date of offense, then measure name similarity by Jaro-Winkler distance. If there is a perfect match on name, we keep only that match. Failing that, we keep matches with a Jaro-Winkler score higher than 0.9 for both first and last name. This is a high threshold but allows some room for spelling and transcription mistakes.

[9] The first seven weeks after the Academy are phase one of field training, the second seven weeks are phase two, the third seven weeks are phase three, and the last three weeks are phase four.

We then identify the senior officer that a recruit is most likely to arrive to a call with during each phase and characterize this officer as the recruit's FTO during that phase. To account for the fact that many officers are assigned to more severe calls, we apply a set of weights equal to the inverse number of senior officers on a given call. The institution of this weighting scheme is that the calls where a recruit arrives with only one other officer (weight=1) provide more information about the identity of their FTO relative to calls where there are many senior officers (weight=$\frac{1}{n}$) on the scene. In our sample, we have a total of 411 recruits and we identify a total of 232 distinct phase 1 FTOs.

The Dallas police department typically requires that FTOs achieve at least the rank of Senior Corporal. We are reasonably confident that we have correctly identified the recruit-FTO pairings in the vast majority of our sample. However, we verify our FTO identification using another dataset documenting overtime pay.[10] In Dallas, each FTO is eligible for overtime pay to compensate for the time spent completing the necessary paperwork to evaluate a recruit after each shift. To check whether our procedure for identifying FTOs is reasonable, we verify that each officer we have flagged as an FTO is observed as receiving overtime during the training period. We find that 390 of the officer-recruit pairs that we have identified as FTO-recruit pairings also appear in the overtime pay dataset while 21 (approximately 5%) are not. Our results are robust to dropping these pairings where we fail to find the FTO in the overtime data.[11] Furthermore, we do not feel that misidentification of these pairings creates any bias in our subsequent results. In particular, we are confident that these are the senior officers that recruits have actually shadowed on the largest number of calls during their initial phase of field training. Thus, these are the senior officers who were most likely to have an impact of a recruit's subsequent policing behavior regardless of whether they were the true administratively assigned FTO. Since our analysis focuses primarily on the impact of the first FTO, we only provide summary statistics related to that pairing.

Police Officers are eligible for promotion to Senior Corporal after three years of service. According to Dallas command staff, most officers who stay with the force for three years should

[10] These data also include comp. time taken instead of overtime.

[11] We estimate coefficients of 0.0002068***, 0.0002038***, and 0.0001519** for our Table 2.4 specifications.

expect a promotion.[12] Although command staff emphasized that there is still some selection in terms of who was allowed to become an FTO, it was not necessarily a position reserved for only highly experienced or exceptionally talented officers. According to our data, the average age of a FTO is 48. This is three years younger than the average age of a patrol officer. FTOs were also generally representative of the whole police force in terms of demographics, but perhaps a bit less diverse. Specifically, 19 percent of FTOs were Hispanic, 16 percent were Black, and 63 percent were White, compared to 20 percent, 23 percent, and 53 percent in the entire force, respectively.

III.B Force Rate Calculation

Next, we assign each of the 411 FTOs-recruit pairs a force rate based on the FTO's propensity to use force in the period prior to being assigned a given recruit. To do so, we estimate pair-specific fixed effects, which represent an FTO's time-invariant propensity to use force on a call for service for a specific recruit. Specifically, we regress an indicator for a call resulting in force on a fixed effect for each recruit-FTO pair using only calls for service answered by the FTO in the period prior to being assigned a given recruit.[13] In estimating this fixed effect, we also control for important call characteristics such as the number of officers on the scene, beat, type of call (priority-by-type) year-by-month, and day of the week-by night fixed effects. The intuition behind this exercise is to create a measure that captures how likely an FTO is to use force in the period prior to being assigned a given recruit and after accounting for the fact that some officers may respond to different types of calls than others. Formally we estimate

$$force_{o(r),c} = \lambda_{o(r)} + \beta_1 X_c + \epsilon_{o(r),c} \tag{2.1}$$

[12] There are four main positions within the Dallas Police Department. Officers begin with the rank of Police Officer and then can advance to Senior Corporal, Sergeant, and finally Lieutenant. Each promotion entails a pay raise.

[13] In practice, this means some FTOs will have more than one fixed effect. Thus, the fixed effect will be unique and estimated separately for each recruit-FTO pairing as opposed to each FTO. These fixed effects will be estimated using the pre-period data relevant to each specific pairing. For example, the fixed effect for a given FTO with a recruit assigned in a later period will leverage more data than the fixed effect for the same FTO assigned to a different recruit in an earlier period. Across different recruits, a given FTO has a remarkable similar propensity to use force. In particular, the correlation coefficient from a comparison of the overall force rate with the pair-specific rate is 0.84.

where $force_{o(r),c}$ is a binary variable equal to one if call c answered by FTO o ends in force and zero otherwise.[14] The vector X_c includes controls that characterize a call for service including indicators for the number of officers on the scene, beat, type of call (priority-by-type), calendar month, and day of the week-by-night. The coefficients of interest $\lambda_{o(r)}$ is a measure of the historic force propensity of FTO o, conditional on call characteristics, prior to being assigned a given recruit r. Since we are stacking the calls for service data for each FTO prior to being paired with each recruit and treating each pairing distinctly, the estimated fixed effects can be interpreted as an FTO's average propensity for using force that is exogenous to the particular recruit. Higher values of $\lambda_o(r)$ indicate that a FTO has historically been more aggressive (i.e. uses force more frequently) while lower values of indicate a FTO is less aggressive (i.e. uses less force). We cluster standard errors on the FTO, rather than the recruit-FTO pair, since some FTOs appear more than once with different recruits.

Since our analysis focuses on cohorts of new recruits that join Dallas PD between July 2014 to December 2018 and our data spans the period from January 2013 to July 2019, the force measure assigned to each FTO-recruit pairing will rely on varying amounts of pre-period calls for service data. In particular, a pairing made in July 2014 will rely on (at most) 1.5 years of pre-period data to calculate the FTO's prior force propensity while a pairing made in December 2018 will rely on (at most) 6 years of pre-period data. In addition to these possible issues related to left truncation, force is also a relatively rare outcome of a call for service with a substantial amount of variation both across and within FTOs. We address both of these concerns by adjusting our estimates of FTO force $\lambda_{o(r)}$ using Empirical Bayes following Weisburst [2022). In particular, we construct a shrinkage factor that attenuates the estimates towards the mean for officers where we observe less pre-period data (due to truncation), observe answering fewer calls for service, or who just have a larger within-officer variance in their propensity to use force. The intuition of applying Empirical Bayes is that the resulting measure will vary principally on FTO force estimates which we have

[14]We denote FTO officer o as a function of recruit r since a given FTO can appear in the sample training multiple recruits. Thus, each force measure is computed using only the pre-period data for a specific recruit.

estimated with the highest degree of confidence.[15]

Formally, we estimate the across officer variance in FTO force, σ_A^2 and a within-officer variance, σ_W^2 [16] Next, we use our two variance measures and the number of observations per officer to estimate a shrinkage factor $\frac{\sigma_A^2}{\sigma_A^2+\frac{\sigma_W^2}{N_{o(r)}}}$. Finally, we construct our final shrunken force rates as

$$\Lambda_{o(r)} = \frac{\sigma_A^2}{\sigma_A^2 + \frac{\sigma_W^2}{N_{o(r)}}} * \lambda_{o(r)} \tag{2.2}$$

where we multiply our shrinkage factor by our original fixed effects. We plot the distribution of police officer force rates for all 411 FTO-recruit pairs in Figure 2.1 for the raw and shrunken measure. As expected, the distribution of the shrunken measure is narrower (has a smaller standard deviation) relative to the unshrunken measure. Values above zero indicate that the field training police officer is more likely to use force relative to the average FTO. A number less than zero indicates that the FTO is less likely to use force relative to the average FTO. For the remainder of our analysis, we will focus on a standardized version of the shrunken FTO force measure (i.e. a z-score) for ease of interpretation.[17]

The distribution of standardized effects is shown in Figure 2.1b. One standard deviation increase in FTO effects is a 0.1706 percentage point which represents a 138 percent increase in average use of force compared to the sample mean of 0.1234 percent. Moving from the FTO that used the least amount of force to the most is an approximate 6 standard deviation increase. Replacing an FTO at the 10th percentile for one at the 90th percentile represents a 338 percent increase in average force.

Finally, we compare FTO force rates to the force rates of other patrol officers in Dallas. To do so, we first construct a force rate for each officer using our entire sample of calls for service. Next,

[15]We can also force FTOs to have the same number of calls before being assigned a recruit (2000 calls). Our main results (Table 4) are similar in magnitude if we make this restriction and re-estimate force rates. The limitation of this method is that we must limit our sample of FTOs.

[16]Formally we calculate within officer residual variance as $\sigma_W^2 = E(\epsilon_{o(r)}^2)$.

[17]We note that our main results are robust to using the unshrunken estimates as well as a number of alternative specifications including a binary indicating high/low force FTOs (see Table B.4).

we shrink and standardized the force rates as described above. Our results are shown in Figure B.2.[18] On average, FTOs use force more, 0.45 standard deviations on average, than the typical non-FTO officer, and more than the average senior Corporal or Sergeant (the ranks most likely to be FTOs). Despite these differences our main takeaway is that there significant overlap between the distributions. While FTOs may be selected on force usage to a certain extent, their propensity to use force does not make them outliers relative to all other patrol officers.

III.C Summary Statistics

We present summary statistics at the recruit level in Table 2.1. As noted, there are 411 recruits in our sample. This translates to roughly 90 new recruits each year. The average recruit is much younger than the average FTO. Most recruits are White (44 percent), 21 percent are Black, and 30 percent are Hispanic. Given the conditional random assignment of recruits to FTOs, we would expect that recruit characteristics shouldn't differ across the type of FTO. Although these summary statistics do not reflect the exact comparison we use in our formal tests of balance where we control for cohort year by division fixed effects, it is worth noting that recruit characteristics look remarkably similar across high and low force FTOs. A t-tests of the difference in recruit characteristics across high and low force FTOs is not statistically significant.[19]

In our main analysis, we evaluate recruit behavior after field training using data on their subsequent calls for service. Summary statistics at the call level are presented in Table 2.2. In our sample, roughly 3.7/100 calls end in any arrest, 2.2/100 end in a misdemeanor arrest and only 1.2/1000 calls end in a use of force. We characterize a call as having involved force or arrest regardless of the specific officer who used force or made the arrest. This conceptual decision was motivated by possible endogenity in terms of the specific officers on the scene of an incident and who actually ends up using force.[20] Our call data also includes other important characteristics that

[18] There are a few (4 percent) very extreme force users in our sample that we drop to create a figure that is easier to "see".

[19] We estimate p-values of 0.5147, 0.3586,0.3399 , 0.1800, 0.1645 for test of difference across means in race, gender and age.

[20] For force incidents, we also require the time on the force report to be between when the first officer arrived and the call was cleared. This sample restriction was made because we suspect some force incidents occur after a suspect is arrested and in police custody.

may impact police officer behavior on the scene. Specifically, we observe the call type, priority (a measure of urgency and severity), location, date and time.

IV Empirical Methods

IV.A Estimation Model

The conditional random assignment of recruits to FTOs provides an ideal setting for investigating how field training shapes a recruit's subsequent policing behavior. We formally explore this question by estimating a model of the form

$$force_{r,c} = \theta_r + \beta_1 \Lambda_{o(r)} + \beta_2 X_c + \epsilon_{r,c} \tag{2.3}$$

Where $force_{r,c}$ is a binary variable equal to one if call c ends in the use of force. Our primary variable of interest $\Lambda_{o(r)}$ represents the propensity of a recruit's FTO to use force in the period prior to their pairing. As discussed, we shrink this measure using Empirical Bayes and standardize it for ease of interpretation. Thus, our coefficient of interest β_1 can be interpreted as the difference in a recruit's likelihood of using force on a given call associated with a one standard deviation increase in their FTO's prior propensity to use force. We control for possible variation across recruits in their initial assignment over time by including θ_r representing a set of 38 Academy cohort year by division fixed effects. To control for variation across calls, we also include X_c representing a vector of call and recruit attributes. In our fully saturated model, this vector includes recruit characteristics (age gender, race), geographic fixed effects (beat), call characteristics (priority, call type), number of officers dispatched, as well as year-by-month, and day of the week-by night fixed effects. Standard errors are clustered at the recruit level.[21]

The model's identifying assumption is that FTO characteristics, primarily prior propensity to use force, are not correlated with recruit characteristics after conditioning on division by cohort year. Therefore, identification relies on the conditional random assignment of recruits to FTOs

[21] We are also robust (i.e. statistically significant at the 5 percent level or less) to two-way clustering by recruit and FTO as well as one-way clustering by recruit or division by cohort year.

within a given division by cohort year. If random assignment of recruits to FTOs did not exist, we would potentially over/under state the impact of an FTO's prior force propensity since this measure might be correlated with other characteristics of a recruit that might also impact the outcome of a call. In other policing agencies where there is not random assignment of recruits to FTOs, it is reasonable to imagine selection across a number of dimensions that could potentially confound the estimates. In the next section, we will empirically demonstrate that FTO characteristics including propensity to use force are uncorrelated with recruit characteristics.

IV.B Research Design

We begin this section by showing that FTO characteristics are not correlated with observable characteristics of their assigned recruit which would potentially confound our main estimates. While we expect this to be true based on discussions with Dallas command staff about FTO assignments in Dallas, we also provide empirical evidence below. To begin, we regress FTO characteristics on recruit characteristics where the unit of observation is a recruit-FTO pair.[22] Each specification includes division by cohort year fixed effects because we believe that FTOs are as-good-as randomly assigned to recruits within cohorts and divisions. Specifically, we investigate whether FTO age, race, and force rate are correlated with recruit age, race, gender, and hire date (measured in years). The results of this test are shown in Table 2.3. Of the 35 coefficients reported, only one is statistically significant at conventional levels.[23] Further, none of the p-values from a joint F-tests are statistically significant at conventional levels.

We also plot the distribution of FTO force rates by recruit characteristics in Figure B.3. Given the random assignment of recruits, we would expect these distributions to be very similar. Indeed, the distributions appear to be very similar and a Kolmogorov-Smirnov test also fails to estimate statistically significant differences across the distributions. These results indicate that recruit characteristics are generally orthogonal to FTO characteristics and are consistent with the institutional

[22]As discussed, a recruit-FTO pair means that there is one observation per recruit but each FTO can be assigned to multiple recruits over the sample period.

[23]Here, we present robust standard errors. However we are robust to clustering at the division by cohort year level or FTO level, i.e. in each three coefficients and none of the F-test p-values in the tables are significant at conventional levels. See Table 2.3 and Table B.2.

background that recruits are as good as randomly assigned to FTOs. Thus, we feel that we are justified in interpretting our results as plausibly causal, i.e. that the coefficient β_1 on the variable $\Lambda_{o(r)}$ from the prior section is the effect of an FTO's propensity to use force on a recruit's subsequent policing behavior.

V Empirical Analysis

V.A Evidence from the Raw Data

We begin by presenting some motivating figures for our analysis. While our formal results always condition on cohort year by division (i.e. the institutional unit where random assignment occurs), we believe it is useful to consider the relationship between a FTO's propensity to use force and their recruit's subsequent use of force in the raw data. In the top panel of Figure 2.2 (a), we plot local average recruit use of force across different FTO force rates. Observations are grouped such that each point includes an equal number of calls. A linear fitted curve is plotted across all force rates. There are two takeaways from this figure. First, we observe both higher and lower force FTO being dispatched to calls. Second, and perhaps most important, there is a clear positive relationship between recruit use of force and their FTO's propensity to use force. The slope of the linear fit suggests that a one standard deviation increase (138 percent) in the FTO's propensity to use force will lead to an increase in recruit force by 0.025 percentage points or approximately 20 percent.[24]

In the bottom panel of Figure 2.2 (b), we again use the raw data to ask whether recruits who are assigned to a high-force FTO are more likely to select into more dangerous calls (ex-ante) that have a high probability of ending in force. If this were true, the data would suggest that the mechanism is driven by selection into more dangerous calls rather than about an officer's willingness to apply force, ceteris paribus. Another interpretation of the predicted results is that if we document a large positive relationship between FTO force and predicted recruit use of force, this could indicate that higher force FTOs lead to more predictable, or reasonable use of force, as opposed to less predictable and more undesirable use of force.

[24] The p-value from this regression is less than 0.001

To assess this alternative explanation, we first regress a recruit's use of force on cohort year and initial assignment fixed effects. We then regress these residuals on every covariate we observe for each call. These include the number of officers on the scene, beat, type of call (priority-by-type) year-by-month, and day of the week-by night. We then use that model to predict the likelihood that force would be used for a given call and add the mean use of force rate to pin down our measure.[25] [26] Although there does appear to be a slight positive correlation between predicted force and FTO force rates, the relationship is much weaker than that shown in the top panel. Given these results and out of an abundance of caution, our preferred specifications include call controls such that we can attempt to rule out the selection channel. Specifically, our fully specified model will include fixed effects for number of officers on the scene, beat, type of call (priority-by-type) year-by-month, and day of the week-by night. We will also expect that the passage of less predicable and potentially less desirable use of force from FTOs to recruits will drive our results.

Taken together, these Figures provide strong supporting evidence that FTOs are an important determinant of recruit use of force. One downside of this graphical analysis is that there is potential for recruit sorting across cohort years and divisions. However in discussion with the Dallas Police Department we believe this type of sorting is limited and not a substantial source of bias. We now turn to our main analysis where we control for cohort year by division fixed effects as well as additional call and recruit characteristics.

V.B Main Results

Next, we present results for the effect of FTOs in Table 2.4. Each specification includes cohort year by division fixed effects, and standard errors are clustered at the recruit level. Our results are also robust to two-way clustering at the recruit badge and FTO level, as well as recruit and division-by-cohort year level.[27] The outcome variable for each column is the proportion of 911 calls that end in force. Column 1 presents our baseline specification where the coefficient on

[25] Intuitively, this produces a linear combination of call characteristics, where the weights are chosen based on the prediction of the likelihood of force being used.

[26] Coefficient on linear fit is 0.0046 percentage points with a p-value of 0.116 . The correlation coefficient for the two terms is 0.0127.

[27] Namely, columns 1 and 2 are significant at the 1 percent level and column 3 is significant at the 5 percent level.

$force_{r,c}$ captures the difference between recruit use of force for recruits assigned to an FTO with one standard deviation higher force propensity (138 percent) . Our results show that recruits with FTOs that use force one standard deviation more are 0.0199 percentage points or 16 percent more likely to use force.[28]

In column 2, we add controls for recruit characteristics (age, gender, race). Given our conditional random assignment and the results in Table B.3 and Figure B.3, we would not expect recruit characteristics to alter our estimate meaningfully. Column 2 shows that recruits with FTOs that use force one standard deviation more are 0.0205 percentage points or 17 percent more likely to use force.

In column 3, we add controls for call characteristics (the same used to predict use of force). Even if recruits are indeed randomly assigned to FTOs, it is possible that the inclusion of call controls could alter our treatment effect. This is because assignment to a high force FTO could cause recruits to work in areas where calls generally tend to be more severe. Another way to interpret column 3, is to think of the inclusion of call controls as controlling for predictable or more desirable use of force. Given the small positive correlation in Figure 2.2, we should expect the magnitude of our estimates to attenuate slightly. Indeed, our estimate in column 3 is about 24 percent smaller but still economically meaningful and statistically significant at the 5 percent level. Even after controlling call characteristics and holding constant this alternative channel, we find a one standard deviation increase (138 percent) in FTO force increases recruit force by 0.015 percentage points or 12 percent. We interpret these and the prior results as providing additional evidence in support of preferred mechanism, i.e. that FTOs transfer information to recruits about the appropriate use of force and that our results are not only driven by more predictable use of force.

Next, we explore whether certain recruits are particularly susceptible to adopting the force behavior of their FTO. In particular, we examine heterogeneity in our main estimates across recruit

[28] In Table B.4 we show results for alternative measure of force rates. Namely, our results are robust to using the unstandardized measure, inverse hyperbolic transformation of our shrunken force measure, and the unshrunk force measure. We also explore using a binary treatment variable, and the results are similar but not always statistically significant at conventional levels.

characteristics like race, gender, and age. Our results are shown in Figure 2.3a where we report the coefficient on FTO force rate. All coefficient estimates, except for that associated with female recruits, are greater than zero. Although there is some variation across these subgroups, the main takeaway of the figure is that all of these subgroups appear to be impacted in the same manner by their FTO's prior propensity to use force.

Finally, we consider the possibility that FTOs with certain characteristics may be better able to transfer their force behavior to recruits. In particular, we test for heterogeneity by FTO characteristics in Figure 2.3b where each coefficient is from a separate regression. Similar to the recruit characteristics plot, every coefficient is greater than zero except for female FTOs. It is also true that younger officers seem to have larger effects in both Figures. Our main takeaway from these figures is that, while there may be some variation across subgroups, the effect of FTO force is prevalent and consistent across nearly all recruits and FTOs. From a policy perspective, this is important because it shows that many different types of recruits could be influenced by reforms to field training or stricter screening of FTOs. Further, back of the envelope calculations show replacing top decile FTOs with bottom decile FTOs would reduce overall force by 5%.

V.C Persistent Effects and Potential Attrition

Understanding the importance of FTOs in terms of force behavior requires considering how long our treatment effect persists. In our setting, this is particularly important given Holz et al. [2020) documents only short-term effects for the same outcome but a different treatment, i.e. the effect of peer injury on use of force.

To consider how our main effects evolve over time, we estimate a model of the form

$$force_{r,c} = \theta_r + \Sigma_{t=0}^{7} \beta_t \Lambda_{o(r)} biannual_t + \beta_2 X_c + \epsilon_{r,c} \tag{2.4}$$

where $biannual$ is a binary variable that takes on a value of 1 t 6 month periods after the end of training. We also add separate fixed effects for $biannual$. All other terms are unchanged from Equation 1 and column 3 controls are included (i.e. call and recruit characteristics). Results are shown in Figure 2.4. For the first two and a half years, every coefficient is greater than zero. After two and a half years the effects appear to attenuate significantly. However, we note that our sample becomes very thin as we examine further time horizons and we are likely relying on less variation in FTO force propensity. That said, it is worth nothing that most officers are promoted to senior corporal after three years of service and themselves either become a detective or FTO.

It is also reasonable to be concerned that our results potentially suffer from selective attrition bias. For example, recruits paired with lower force FTOs might be more likely to be terminated or take assignments where they no longer respond to calls for service. With respect to termination, it is highly unlikely that our results are driven by attrition along this dimension as only three recruits leave the Dallas Police Department in the first three years after training. With respect to recruits accepting alternate assignments where they no longer respond to calls for service, we find only 1% of sample exit these data in months 1-23 but an additional 9% exit in month 24 and 13% exit in months 25-30. We address this potential concern by first asking whether FTO force rates are correlated with the last date we observe a recruit in the calls for service data.[29] Regressing each recruit's exit date on FTO force, we do not find evidence that FTO force predicts when recruits stop responding to calls for service in our data. Next, we alternatively address this potential concern by limiting our estimation sample to the calls for service data occurring in months 1-23 when there is virtually no attrition in our data. Using this alternate sample, we find very similar results to Table B.5. Given both of these results, we believe that we can confidently put aside potential concerns of attrition bias driving our results.

V.D Randomization Inference

In this section, we provide a robustness test focusing on the calculation of standard errors in our main results, i.e. columns 1-3 of Table 2.4. In our main estimates, we follow standard approaches by clustering our standard errors at the recruit level and note that we are also robust to two-way clustering at the recruit and division by year level (Bertrand et al., 2004). The concern motivating

[29] Specifically, we regress the last date we observe a recruit in the calls data on FTO force rate using the specifications shown in column 1, 2, and 3 of Table 2.4. None of the coefficients on FTO force were statistically significant at conventional levels (p-values = 0.364, 0.166, 0.147).

the robustness test in this section is that our outcome variable (force by call) is a rare event, occurring in only 0.106 percent of calls from our sample. In cases where an outcome variable is a rare event, standard asymptotic assumptions related to the distribution of point estimates and associated standard errors may be inappropriate. Here, we use randomization inference to construct an empirical distribution of point estimates and reassess the validity of the hypothesis tests conducted for our primary set of estimates.

As discussed in Efron [2004), randomization inference is most appropriate to non-experimental settings when researchers are able to replicate the data generating process of the observed data. In our institutional setting, recruits from a given Academy cohort are randomly assigned to FTOs by command staff within their respective division. As discussed previously, our balancing tests support that this source of variation is as-good-as random. Thus, our randomization procedure attempts to replicate this variation in constructing an empirical distribution of point estimates and associated standard errors. For each recruit in our sample, we randomly draw an FTO from the set of eligible officers we observe working in the recruit's respective division. [30] As with our main estimates, we next construct an estimate of FTO force propensity using calls answered by the randomized FTO in the period prior to being assigned the particular recruit. We then shrink that estimate using the Empirical Bayes procedure described in the methods section and standardize the shrunken measure by subtracting the mean and dividing by the standard deviation. Using the randomized FTO's propensity to use force in the pre-period as the primary independent variable, we estimate the model from columns 1-3 of Table 2.4. In order to obtain p-values for a two-sided hypothesis test, we replicated this procedure 1,000 times and calculate the share of the simulations when the t-statistic exceeds the absolute value of the t-statistics from Table 2.4.

For the models corresponding to columns 1 and 2 of table 2.4, we obtain p-values for a two-sided hypothesis test using randomization inference of 0.01 and 0.013 respectively. For illustrative purposes, we also plot the distribution of t-statistics obtained from our randomization procedure in

[30]We consider an officer as eligible for being a given recruit's FTO if they are observed answering at least one call for service in the same division within 30 days of the recruit's first day assigned to patrol. We also only consider officers as eligible to be an FTO if they have a rank of police officer or higher though we note that we are robust to imposing a more stringent rank requirement of senior corporal or above.

figure 2.5 corresponding to the fully specified model in column 3 of table 2.4. In our randomization procedure, we find that eighty of the simulations result in a t statistic more extreme than 2.137. Thus, the associated p-value obtained through randomization inference is 0.08. We interpret these results as providing additional evidence indicating that the inference in our main results is not entirely driven by potential inference issues associated incorrect asymptotic assumptions or rare-event bias.

VI Mechanism and Recruit Arrests

VI.A Alternative Mechanisms

Until this point, we have interpreted our main results an FTO transferring information about the appropriate threshold for applying force to their recruit, i.e. an FTO teaching a recruit to behave more/less aggressively. In this section, we will provide additional evidence in support of that particular explanation of the underlying mechanism by convincingly ruling out six alternative explanations. These alternative explanations include: (1) omitted FTO characteristics are correlated with FTO force and a recruit's subsequent force; (2) whether our main results are driven by more active forms of policing; (3) whether recruits are more likely to be dispatched to calls for service with their FTO even after they complete training; (4) whether recruits paired with a high-force FTO are actually witnessing force during the training period; (5) differential behavior in terms of reporting force; and (6) whether our results are driven by under-reporting of less severe force incidents. Across all of these additional estimates, we find very little evidence of any mechanism other than knowledge transfer from the FTO to the recruit about the appropriate use of force.

First, we explore whether FTO demographics are predictive of force rates. In Figure B.1, we plot the distribution of force rates by FTO characteristics. On average white FTOs have higher force rates than Hispanic and Black FTOs but none of the distributions are statistically different from each other.[31] Figure B.1b shows that female FTOs are nearly half a standard deviation more likely to use force than males and these two distributions are statistically different from each other

[31] The average Black FTO has a force rate that is about 1/10 of a standard deviation higher than the average White FTO. We also note that there are few Black and Hispanic FTOs constituting only 70 and 68 recruit-FTO pairs respectively

(K-Smirnov p-value = 0.001). Figure B.1a also shows that younger FTOs (i.e. less than the mean age of 50) use force 0.15 standard deviations more frequently than older FTOs and the two distributions are statistically different from each other (K-Smirnov p-value = 0.09).[32] Finally, we regress our force measure on FTO race, age, and gender. This regression yields an R-squared of 0.05 and an F-test p-value of 0.02. Since we observe that force rates are correlated with other FTO characteristics, we now formally control for these measures and re-estimate the main results from column 3 of Table 2.4 (shown again in column 1) in columns 2, 3, and 4 of Table 2.6. Overall, these estimates are at least as large as our main results and statistically significant at conventional levels.

Second, we explore whether FTO force is capturing proactive policing, as opposed to a lower threshold for applying force. We construct a set of additional measures that capture other aspects of FTO behavior that we believe are associated with proactive policing. In particular, we construct measures of average FTO arrest (misdemeanor and overall) as well as response times and time spent on a call.[33] We begin by presenting the correlation between FTO force, and other FTO rates in Table 2.5. Column 1 in Panel A reports that a one standard deviation increase in FTO force propensity is associated with a 0.58 standard deviation increase in arrest propensity. We also find similar effects for other types of arrests (columns 2-5). Column 1 in Panel B considers the correlation between FTO force rates, response time, and time spent on a call. Although we find no statistically significant correlation with response time, we find that FTO's with higher force rates tend to spend less time on a call.This result is in line with our conversations with Dallas FTOs, who claim more engaged, or less "lazy", officers do not loiter at the end of calls but instead quickly respond to other calls. Since we observe that FTO force is correlated with other measures that capture a more active form of policing, we now formally control for arrest rates, misdemeanor arrest rates, response time and time on call in columns 5-8 of 2.6. Overall, these estimates are similar in size and significance to our main results shown again in column 1.

[32] We also regress our force rate on a FTO age and find that 10 years of age translate to a 0.11 standard deviation decrease in an officer's force rate, however this estimate is not significant at conventional levels.

[33] Specifically, we estimate Equation (1) using arrest, misdemeanor arrest, response time and time on calls as the outcome and shrink our estimates according to Equation (2).

We include all FTO characteristics (columns 2-4) as well as other measures of proactive policing (columns 5-8) in column 9 of 2.6. Our estimate is larger in magnitude than column 1 and statistically significant at the one percent level. Together, these results illustrate that FTOs transferring information about the appropriate threshold for applying force, as opposed to other FTO characteristics or proactive policing, is the most likely explanation for our results.

Third, we ask whether our results are driven by who a recruit is dispatched with after training. First, we consider times when a recruit is dispatched with their FTO after training is complete. There are few calls where we observe FTO-recruit pairs dispatched together (16,562). If we drop those calls from our sample, our Table 2.4 column 3 (full controls) estimate is a similar magnitude and is statistically significant at the 5% level. We do not believe our results are driven by the continued pairing of recruits and FTOs. Next, we look at other partners (the officer a recruit is observed the most with after training) in Table 2.7. If we control for partner force rate, measured across the entire sample, and other partner characteristics (age, gender, race), our results shrink some, but remain similar to the estimates in Table 2.4 and significant at conventional levels. We note that this approach is not our preferred specification as who a recruit chooses to work could be affected by treatment. Said another way, although including partner characteristics as controls is technically an over-control, we believe it illustrates that partners do not fully explain our results.

Fourth, we ask whether recruits paired with more aggressive FTOs might be more likely to experience force during their training periods. This early exposure to a force event could drive our results. Ninety-eight recruits (24%) experienced a force incident during training. To investigate this explanation, we allow our effect to vary by whether a recruit experienced force during their training. Figure 2.6 shows that results are similar no matter if a recruit experienced force during their training period. Within each specification (following specifications in Table 2.4), coefficients are not statistically different from each other. If anything, results are larger for recruits who did not experience force during their training. The results of this section show that results are driven by FTO force, not other characteristics of FTO policing style, other attributes of a recruit's training period, or peers after training.

Fifth, we explore whether our results are potentially driven by differential reporting patterns amongst high force FTOs that are transmitted to recruits. In conversations with both police officers and command staff at Dallas PD, we asked Dallas Police Department officers about force reporting norms within the department. Every member of the Dallas PD we spoke to reiterated that all force incidents are recorded in BlueTeams and that under reporting was unlikely because the department conducts frequent audits of reports and bodyworn/dashboard camera footage. If an officer were to be caught engaging in unreported force, they would face serious repercussions.[34] To explore differences in report-writing behavior, we develop three measures based on the section of incident reports completed by the responding officer. [35] Following the procedure described in III.C, we calculate measures the capture the total number of characters written in incident reports by the FTO as well as number of distinct words and a variable capturing not having entered anything. In panel C of Table 2.5, we show that our proxies for reporting writing are not correlated with FTO force. In Table 2.9, we repeat the three columns presented in Table 2.4 but with the addition of our reporting controls. Across each column each of the estimates are similar in magnitude and significance to Table 2.4.[36]

Finally and relevant to the prior discussion, we again explore under-reporting by focusing on a subset of the most "reportable" force incidents, i.e. those where a weapon was discharged.[37] A limitation of our data set is that the information on the type of force used is missing for 70% of our linked force incidents. For the following analysis we drop those force incidents with missing information. When replace our outcome variable with one for *only* the most severe force incidents, our results are remarkably similar (see Table B.6).[38] Further our results are not driven by events

[34]The police department's General Orders reiterate the auditing of reports and bodyworn/dashboard camera footage stating "Supervisors will conduct random BWC reviews/audits of personnel assigned to them" (Dallas Police Department [2021]). It is also worth noting that, even if an officer inappropriately uses force on a citizen, the incentives are such that they are actually better off documenting the incident as opposed to potentially receiving a complaint about an unreported incident.

[35]In particular, we rely on the field "Modus Operandi (MO)" which, to our knowledge, is an open-ended text field that is completed by the officer taking the incident report.

[36]There are 10 FTOs that we cannot link to incident reports. Therefore we have approximately 10 thousand fewer observations in Table 2.9 relative to our primary estimates.

[37]Specifically we flag force incidents where a firearm, taser, spray, or pepperball was used.

[38]Given how rare our outcome measure is, we shift from beat to sector fixed effects, and estimate fixed effects for whether a recruit is above average age (instead of a fixed effect for each year).

with only one officers, where there is arguably less incentive to report.[39]

VI.B Arrest Results

Thurs far, our paper has demonstrated that recruits assigned to a high-force FTO are more likely to use force later in their careers. One limitation of most existing work on police use of force, is that it is arguably impossible for the researcher to determine which interactions or calls for service warrant use of force (a“good" application of force) and which do not (a “bad" application of force). Put differently, there are clearly some dangerous incidents when we want to empower police to use force in the name of preserving public safety. In an effort to explore whether high-force FTOs are engaged in other behaviors that might be considered "good" vs. "bad" from a social or policy perspect, we consider four additional measures in Table 2.8. Panel A presents results for arrests that are arguably less desirable from a social/policy perspective, and Panel B presents results for more serious and higher quality arrests. Our results follow the columns presented in Table 2.4 where we demonstrate that an increase of one standard deviation in FTO force was associated with a 3-5% increase in the likelihood of a misdemeanor arrest and a 3-7% increase in the likelihood of an unfiled arrests. Unfiled arrest are arrests that are made, but not formally filed at the District Attorney's office. This typically happens when an arrest is made for frivolous reasons or without proper evidence or supporting documentation. In Panel B, we find that recruits with higher force FTOs do not make more felony arrests or filed arrests. Together, Panels A and B provide evidence that high-force FTOs not only lead to high-force recruits but that they also produce recruits that make more petty arrests and unfiled "low quality" arrests but not more serious "high quality" felony arrests. In short, we find very little evidence in support of the idea that high force is associated with the types of behaviors we might characterize as "good" policing.

VII Conclusion

This paper estimates the effects of high force FTOs on recruit use of force. We compare recruits quasi-randomly assigned FTOs with higher historical force rates to those with lower force FTOs.

[39]Our results are still statistically significant at conventional levels and larger when we drop calls with only one officer.

Our results show FTOs are an important determinant of subsequent recruit force; a one standard deviation increase in FTO force increases recruit force by 12 percent. FTO effects also persist two and a half years after completing training. We also show that our results are consistent with the story of less desirable policing qualities alone being transferred from FTO to recruit. Given the broad support for reforms to police officer training, the wide availability of alternative FTOs, and the relative ease of switching FTOs, we believe our findings are consistent with FTOs being a viable avenue for reducing aggressive policing.

VIII References

M. Anelli and G. Peri. The effects of high school peers' gender on college major, college performance and income. *The Economic Journal*, 129(618):553–602, 2019.

B. A. Ba, D. Knox, J. Mummolo, and R. Rivera. The role of officer race and gender in police-civilian interactions in Chicago. *Science*, 371(6530):696–702, 2021.

M. Bertrand and A. Schoar. Managing with style: The effect of managers on firm policies. *The Quarterly Journal of Economics*, 118(4):1169–1208, 2003.

M. Bertrand, E. Duflo, and S. Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1):249–275, 2004.

R. Bifulco, J. M. Fletcher, and S. L. Ross. The effect of classmate characteristics on post-secondary outcomes: Evidence from the add health. *American Economic Journal: Economic Policy*, 3(1): 25–53, 2011.

S. B. Billings and K. T. Schnepel. Hanging out with the usual suspects: Neighborhood peer effects and recidivism. *Journal of Human Resources*, 2020.

S. B. Billings, D. J. Deming, and S. L. Ross. Partners in crime. *American Economic Journal: Applied Economics*, 11(1):126–150, 2019.

N. Bloom and J. Van Reenen. Measuring and explaining management practices across firms and countries. *The Quarterly Journal of Economics*, 122(4):1351–1408, 2007.

G. Bruinsma and D. Weisburd. *Encyclopedia of criminology and criminal justice*. Springer Reference, 2014.

Bureau of Justice Statistics. State and Local Law Enforcement Training Academies, 2018 – Statistical Tables. `https://bjs.ojp.gov/sites/g/files/xyckuh236/files/media/document/slleta18st.pdf`, 2018.

S. E. Carrell, B. I. Sacerdote, and J. E. West. From natural variation to optimal policy? the importance of endogenous peer group formation. *Econometrica*, 81(3):855–882, 2013.

J. B. Chan. *Changing police culture: Policing in a multicultural society*. Cambridge University Press, 1997.

J. P. Crank. Understanding police culture. 2004.

Dallas Police Department. Dallas Police Department General Orders. `https://dallaspolice.net/resources/Shared%20Documents/General-Orders.pdf`, 2021.

B. Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004.

A. Fenizia. Managers and productivity in the public sector. *Working Paper*, 2021.

R. G. Fryer Jr. An empirical analysis of racial differences in police use of force. *Journal of Political Economy*, 127(3):1210–1261, 2019.

Gallup Panel. Americans' Views of the Need for Changes in Policing. `https://news.gallup.com/poll/315962/americans-say-policing-needs-major-changes.aspx`, 2020.

R. M. Getty, J. L. Worrall, and R. G. Morris. How far from the tree does the apple fall? Field training officers, their trainees, and allegations of misconduct. *Crime & Delinquency*, 62(6): 821–839, 2014.

M. Giorcelli. The long-term effects of management and technology transfers. *American Economic Review*, 109(1):121–52, 2019.

E. L. Glaeser, B. Sacerdote, and J. A. Scheinkman. Crime and social interactions. *Quarterly Journal of Economics*, 111(2):507–548, 1996.

M. Hoekstra and C. Sloan. Does race matter for police use of force? evidence from 911 calls. *American Economic Review*, forthcoming, 2020.

J. E. Holz, R. G. Rivera, and B. A. Ba. Peer effects in police use of force. *Working Paper*, 2020.

Ipsos. USA Today/Ipsos Crime and Safety Poll. `https://www.ipsos.com/sites/default/files/ct/news/documents/2021-07/Topline-USAT-Crime-and-Safety-070821.pdf`, 2021.

H. Johnson, S. Ross, S. Mello, and M. Ross. Experience, formalized training, and police discrimination. *Working Paper*, 2021.

E. P. Lazear, K. L. Shaw, and C. T. Stanton. The value of bosses. *Journal of Labor Economics*, 33 (4):823–861, 2015.

A. Mas and E. Moretti. Peers at work. *American Economic Review*, 99(1):112–45, 2009.

K. McLean, S. E. Wolfe, J. Rojek, G. P. Alpert, and M. R. Smith. Randomized controlled trial of social interaction police training. *Criminology & Public Policy*, 19(3):805–832, 2020.

F. X. Murphy. Does increased exposure to peers with adverse characteristics reduce workplace performance? Evidence from a natural experiment in the US army. *Journal of Labor Economics*, 37(2):435–466, 2019.

E. Owens, D. Weisburd, K. L. Amendola, and G. P. Alpert. Can you build a better cop? experimental evidence on supervision, training, and policing in the community. *Criminology & Public Policy*, 17(1):41–87, 2018.

E. Paoline III. The myth of a monolithic police culture. *Demystifying crime and criminal justice*, pages 81–88, 2006.

E. A. Paoline III. Taking stock: Toward a richer understanding of police culture. *Journal of criminal justice*, 31(3):199–214, 2003.

E. A. Paoline III and J. M. Gau. Police occupational culture: Testing the monolithic model. *Justice Quarterly*, 35(4):670–698, 2018.

Pew Research Center. Majority of public favors giving civilians the power to sue police officers for misconduct. `https://www.pewresearch.org/politics/2020/07/09/majority-of-public-favors-giving-civilians-the-power-to-sue-police-office` 2020.

President's Task Force on 21st Century Policing. Final report of the president's task force on 21st century policing. 2015.

N. Rakich. How americans feel about 'defunding the police'. `https://fivethirtyeight.com/features/americans-like-the-ideas-behind-defunding-the-police-more-than-the-slogan` 2020.

R. Rivera. The Effect of Minority Peers on Future Arrest Quantity and Quality. *Available at SSRN 4067011*, 2022.

D. P. Rosenbaum and D. S. Lawrence. Teaching procedural justice and communication skills during police–community encounters: Results of a randomized control trial with police recruits. *Journal of Experimental Criminology*, 13(3):293–319, 2017.

B. Sacerdote. Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly Journal of Economics*, 116(2):681–704, 2001.

B. Schaefer and T. Hughes. Honing interpersonal necessary tactics (hint): An evaluation of procedural justice training. *Louisville, KY: Southern Police Institute, University of Louisville*, 2016.

W. G. Skogan, M. Van Craen, and C. Hennessy. Training police for procedural justice. *Journal of experimental criminology*, 11(3):319–334, 2015.

J. H. Skolnick. *Justice without trial: Law enforcement in democratic society*. 1966.

Warners, Ronald. The field training experience: Perspectives of field training officers and trainees. `https://www.policechiefmagazine.org/the-field-training-experience-perspectives-of-field-training-officers-and` 2020.

E. K. Weisburst. Police use of force as an extension of arrests: Examining disparities across civilian and officer race. In *AEA Papers and Proceedings*, volume 109, pages 152–56, 2019.

E. K. Weisburst. "whose help is on the way?" the importance of individual police officers in law enforcement outcomes. *Journal of Human Resources*, pages 0720–11019R2, 2022.

J. West. Learning by doing in law enforcement. *Working Paper*, 2019.

W. A. Westley. *Violence and the police: A sociological study of law, custom, and morality*, volume 28. MIT press Cambridge, MA, 1970.

L. Wheller, P. Quinton, A. Fildes, and A. Mills. The greater manchester police procedural justice training experiment. *Coventry, UK: College of Policing*, 2013.

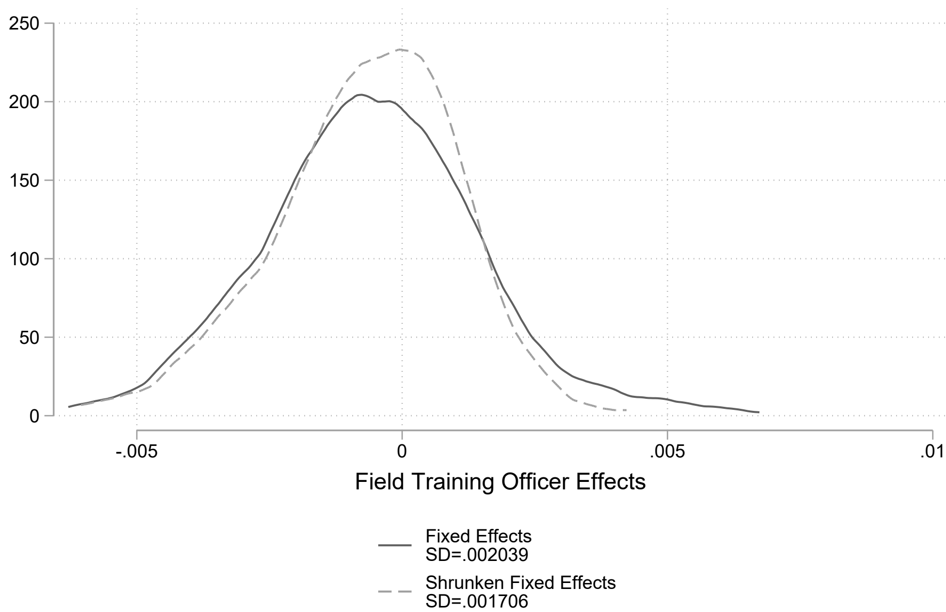
D. Whitmore. Resource and peer impacts on girls' academic achievement: Evidence from a randomized experiment. *American Economic Review*, 95(2):199–203, 2005.

R. H. Woody. The police culture: Research implications for psychological services. *Professional Psychology: Research and Practice*, 36(5):525, 2005.

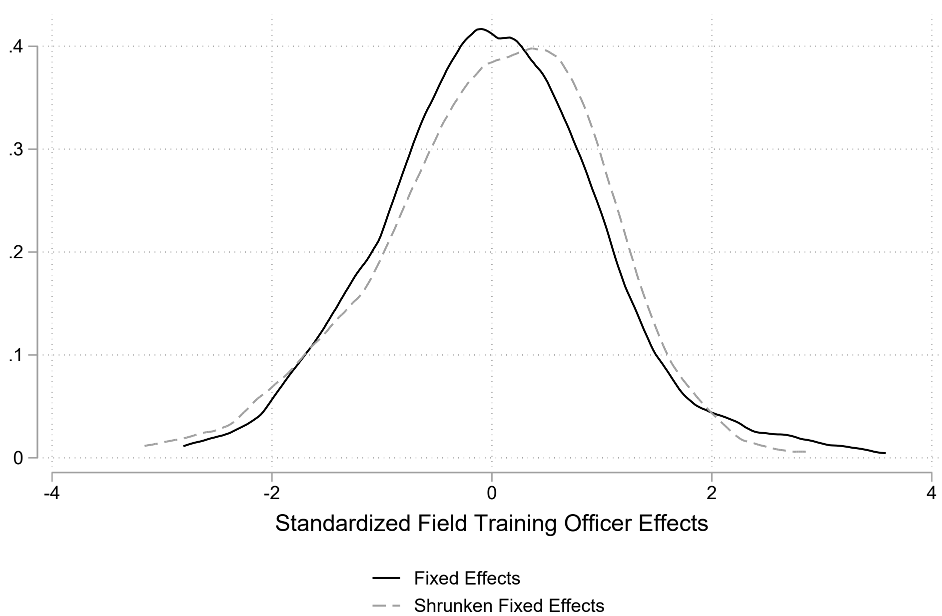
IX Figures

Figure 2.1: Density of Field Training Officer Propensity to Use Force

(a) Field Training Officer Effects



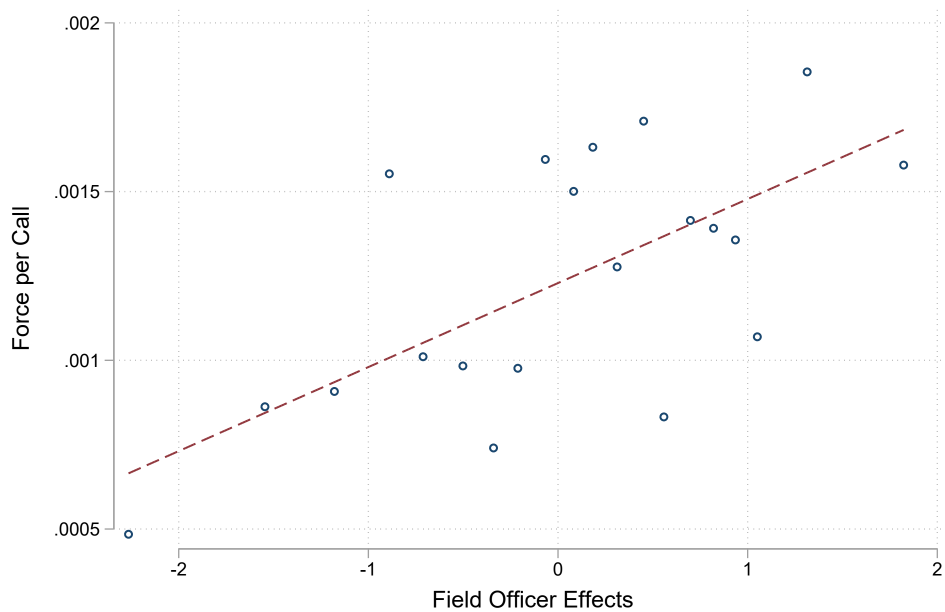
(b) Standardized Field Training Officer Effects



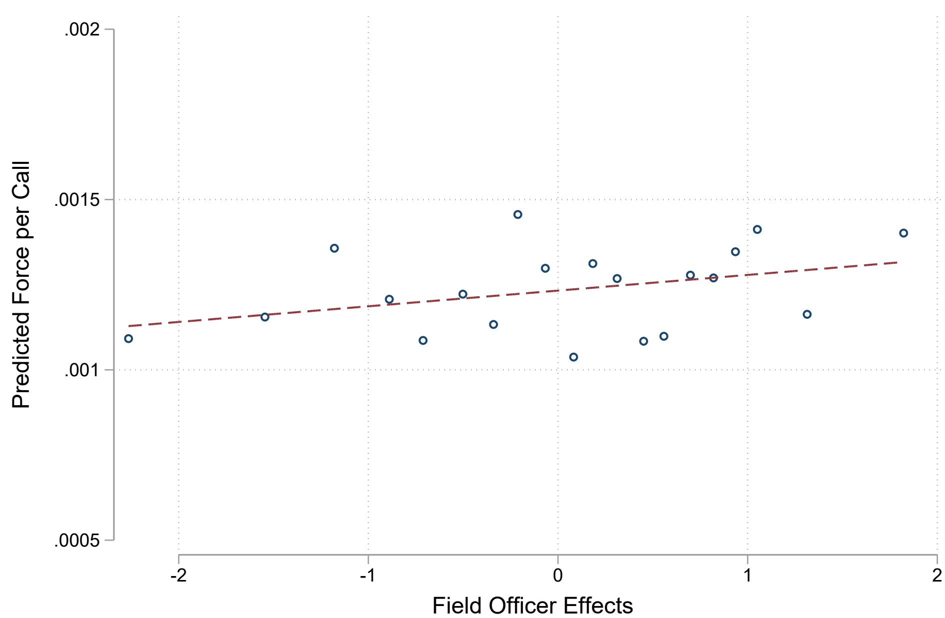
Notes: These figures plot the distribution of field training officer fixed effects. Fixed effects are calculated after accounting for the number of officers on the scene, beat, type of call (priority-by-type) year-by-month, and day of the week-by night fixed effects.

Figure 2.2: Recruit Actual Force and Predicted Force by Field Training Officer Effects

(a) Use of Force

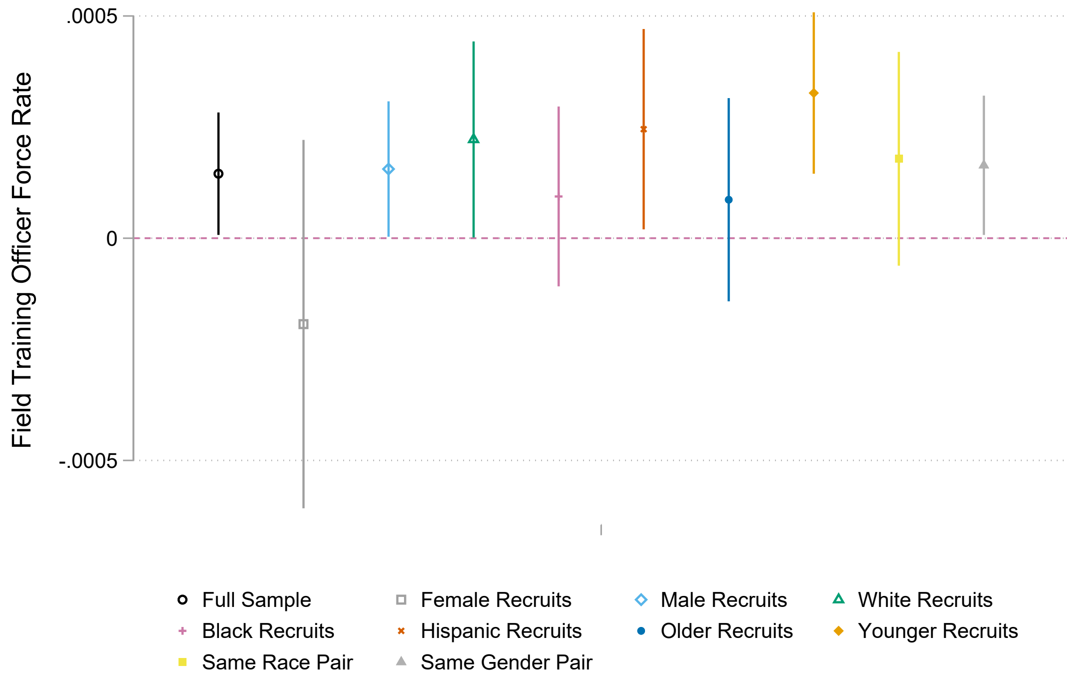


(b) Predicted Use of Force

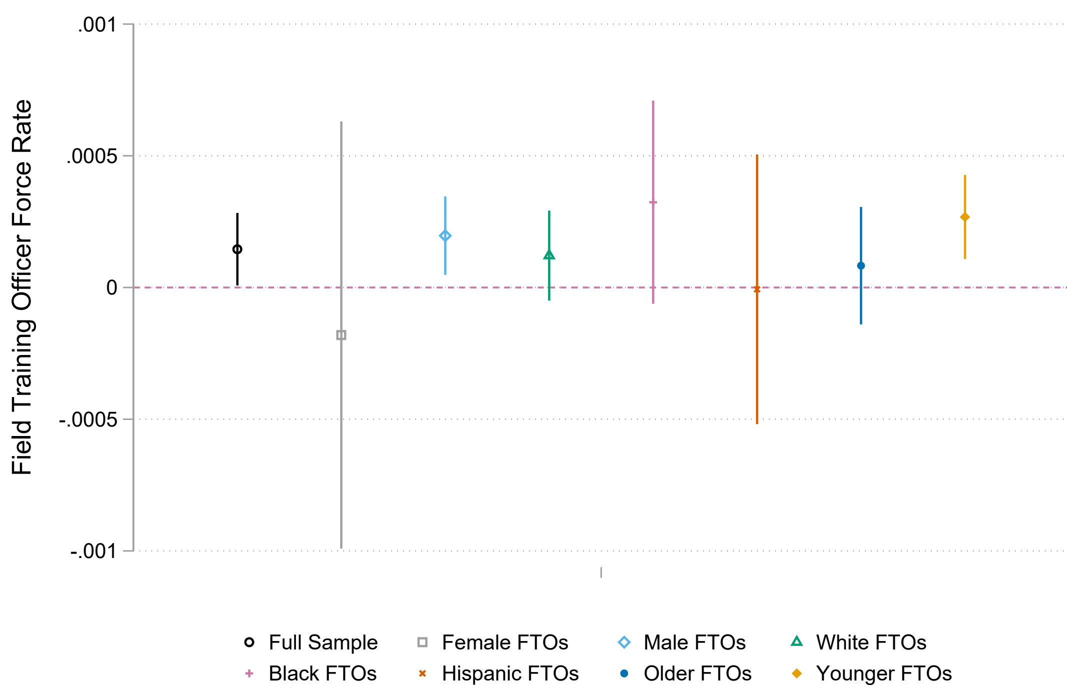


Notes: In Panel (a) we plot use of force. In Panel (b) we plot predicted use of force. The fitted line is a linear fit across all use of force rates. Observations are grouped so that each point includes an equal number of observations. Use of force is predicted using the number of officers on the scene, beat, type of call (priority-by-type) year-by-month, and day of the week-by night fixed effect.

Figure 2.3: The Effect of Field Training Officers on Force by Recruit and FTO Subgroups



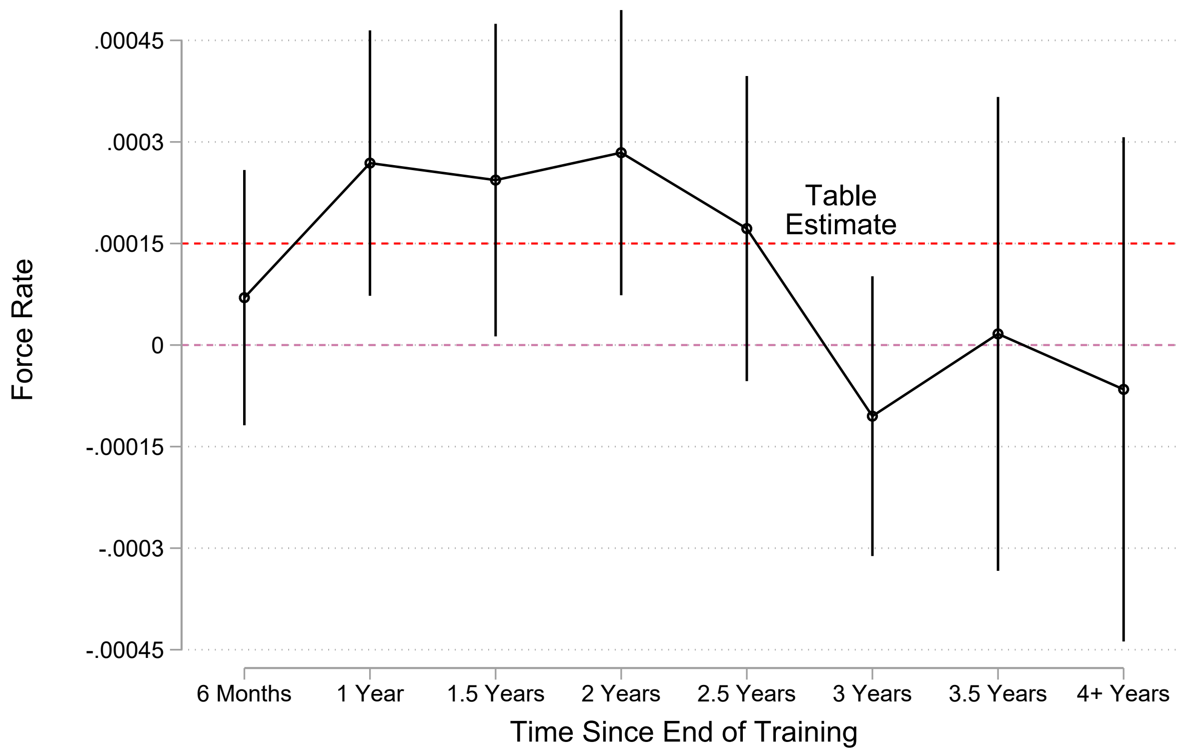
(a) Recruit Subgroups



(b) FTO Subgroups

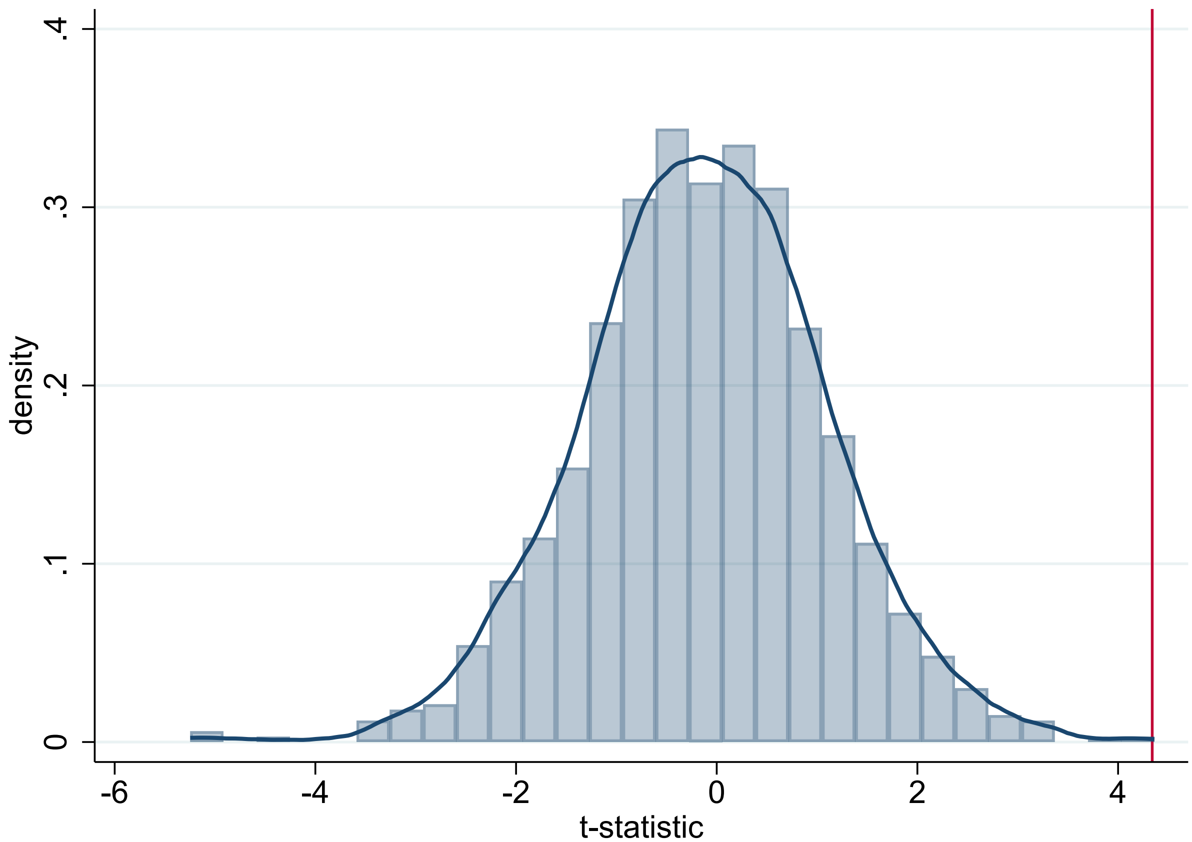
Notes: This figure reports the effect of field training officer force rates by recruit and FTO subgroup. Each coefficient is from a separate regression. Standard errors are clustered at the recruit level.

Figure 2.4: The Effect of Field Training Officers on Force Over Time



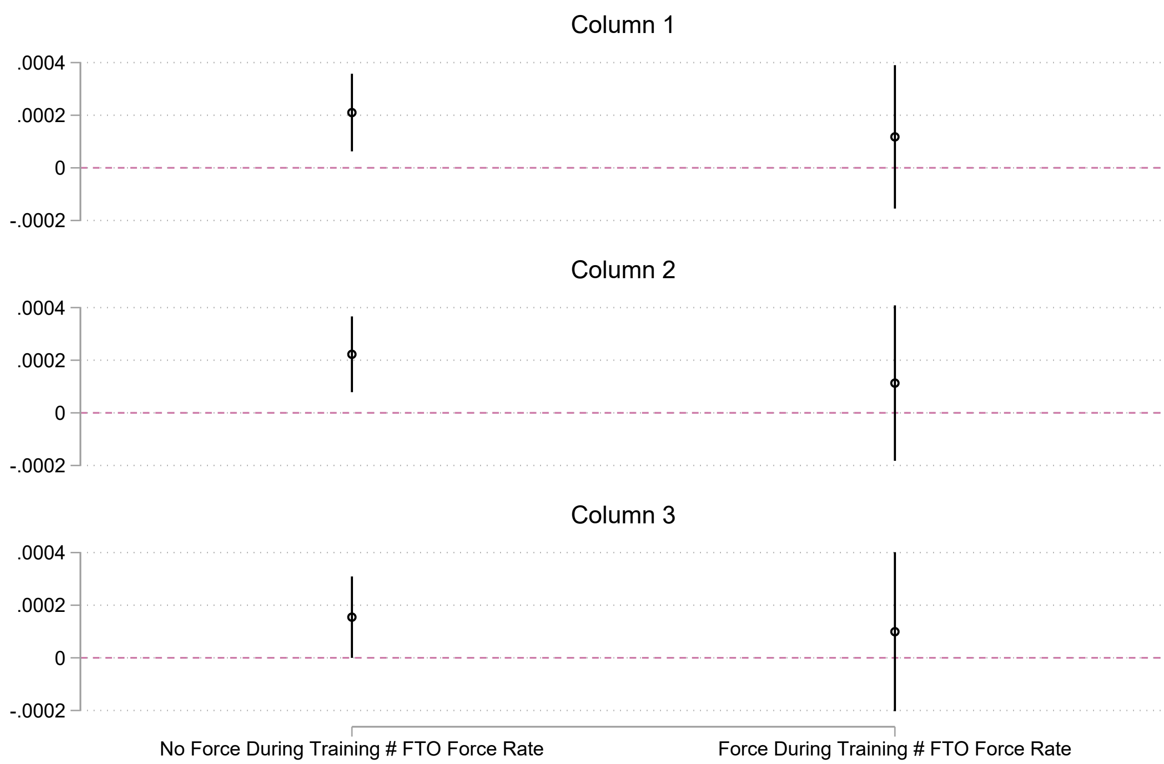
Notes: This figure reports the effect of field training officer force rates over time. Sample includes all officers observed with three full years of data. The red line marks the estimate from column 3 in Table 2.4. Standard errors are clustered at the recruit level.

Figure 2.5: Empirical Distribution of T-Statistics from Randomization Inference



Notes: This figure reports the empirical distribution of t-statistics from estimating the regression shown in column 3 of Table 2.4 by using a 1,000 randomized simulations of the data generating process. Eighty of the simulations resulted in a t-statistic more extreme than our baseline estimate of 2.137 indicating a p-value of less than 0.08.

Figure 2.6: The Effect of Field Training Officers on Force by Recruit Force Experience



Notes: This figure reports the effect of field training officer force rates for recruits that did and did not experience a force incident during their training period. Columns refer to the specifications used in Table 2.4.

X Tables

Table 2.1: Officer-level Summary Statistics

	(1) Entire Sample	(2) High Force Trainer	(3) Low Force Trainer
White	0.438	0.418	0.451
	(0.497)	(0.495)	(0.499)
Black	0.207	0.184	0.221
	(0.406)	(0.388)	(0.416)
Hispanic	0.302	0.329	0.285
	(0.460)	(0.471)	(0.452)
Female	0.178	0.146	0.198
	(0.383)	(0.354)	(0.399)
Age	35.84	36.32	35.55
	(5.466)	(5.948)	(5.132)
Observations	411	158	253

Table 2.2: Call Level Summary Statistics

	(1) Entire Sample	(2) High Force Trainer	(3) Low Force Trainer
Force	0.00123	0.00141	0.00112
	(0.0351)	(0.0375)	(0.0334)
All Arrests	0.0370	0.0374	0.0367
	(0.189)	(0.190)	(0.188)
Misd. Arrest	0.0215	0.0221	0.0211
	(0.145)	(0.147)	(0.144)
Felony Arrest	0.00755	0.00755	0.00754
	(0.0865)	(0.0866)	(0.0865)
Filed Arrest	0.0148	0.0149	0.0148
	(0.121)	(0.121)	(0.121)
Unfiled Arrest	0.0143	0.0149	0.0139
	(0.119)	(0.121)	(0.117)
Observations	1085020	421901	663119

Table 2.3: Balance Test: Correlation between Recruit and Field Training Officer Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
				FTO Characteristics			
	Age	Female	White	Black	Hispanic	Hire Date	Force Rate
Recruit Chars							
White	-1.0720	-0.1124	-0.1555	-0.0052	0.0952	281.5651	-0.3147
	(1.8062)	(0.0879)	(0.1044)	(0.0831)	(0.0728)	(758.3525)	(0.2124)
Black	-1.4251	-0.0514	-0.0800	0.0090	0.0061	334.1880	-0.4693**
	(1.9980)	(0.0971)	(0.1141)	(0.0921)	(0.0793)	(797.1520)	(0.2370)
Hispanic	-1.3049	-0.0561	-0.1512	0.0158	0.0730	418.5266	-0.3657
	(1.8585)	(0.0916)	(0.1088)	(0.0888)	(0.0754)	(765.9610)	(0.2304)
Female	-0.1773	0.0293	-0.0923	0.0502	0.0560	289.7770	-0.1121
	(1.0851)	(0.0472)	(0.0648)	(0.0523)	(0.0537)	(377.1587)	(0.1179)
Age	0.0502	-0.0000	0.0027	0.0003	0.0004	10.6934	0.0051
	(0.0894)	(0.0035)	(0.0051)	(0.0041)	(0.0040)	(30.9955)	(0.0098)
Observations	411	411	411	411	411	411	411
Div-x-Cohort FE	X	X	X	X	X	X	X
Outcome Mean	49.45	0.129	0.628	0.170	0.165	15298	-1.51e-09
F-Test P-Value	0.972	0.482	0.357	0.942	0.359	0.941	0.391

Notes: Robust standard errors are presented.

Table 2.4: The Effect of High Force Field Training Officers on Recruit Use of Force

	(1) Force	(2) Force	(3) Force
FTO Force Rate	0.000199***	0.000205***	0.000150**
	(0.0000663)	(0.0000651)	(0.0000702)
Observations	1085020	1085020	1085020
Outcome Mean	0.00123	0.00123	0.00123
Assigned Div by Cohort FE	Y	Y	Y
Recruit Characteristics	-	Y	Y
Call Controls	-	-	Y

Standard errors in parentheses
* $p < .1$, ** $p < .05$, *** $p < .01$

Notes: Standard errors are clustered at the recruit officer level. Column 2 adds controls for recruit characteristics (age, gender, race). We add call characteristics fixed effects (number of officers on the scene, beat, type of call—priority-by-type,year-by-month, and day of the week-by night) in column 3. A one standard deviation increase in FTO use of force is a 138 percent increase.

Table 2.5: Correlation between FTO Force Rate and Other FTO Behavior

	Overall Arrest Rate	Filed Arrest Rate	Unfiled Arrest Rate	Misd. Arrest Rate	Felony Arrest Rate
Panel A: FTO Arrest					
Force Rate	0.584^{***}	0.596^{***}	0.453^{***}	0.498^{***}	0.625^{***}
	(0.0551)	(0.0516)	(0.0557)	(0.0541)	(0.0533)
Observations	411	411	411	411	411
	Response Rate	Time on Call			
Panel B: FTO Timing					
Force Rate	-0.0299	-0.168^{***}			
	(0.0743)	(0.0473)			
Observations	411	411			
	Num. Characters Rate	Num. Words Rate	Write Nothing Rate		
Panel C: FTO Reporting					
Force Rate	-0.119^{*}	-0.106	0.0317		
	(0.0698)	(0.0670)	(0.0701)		
Observations	401	401	401		
Div-X-Cohort FE	Y	Y	Y		

Standard errors in parentheses
* $p < .1$, ** $p < .05$, *** $p < .01$

Notes: All variables are standardized. Standard errors are clustered at the field training officer level. There are 10 FTO-recruit pairs that we cannot match to incident reports (in Panel C we have 401 FTO-recruit pairs instead of 411). Response time is the number of hours between arrival time and assigned time. Time on call is the number of hours between the time an officer was enroute and when the call was cleared. A one standard deviation increase in FTO use of force is a 138 percent increase.

Table 2.6: Mechanisms: The Effect of High Force Field Training Officers on Recruit Use of Force

	(1) Force	(2) Force	(3) Force	(4) Force	(5) Force	(6) Force	(7) Force	(8) Force	(9) Force
FTO Force Rate	0.000150**	0.000158**	0.000159**	0.000237***	0.000255***	0.000217***	0.000151**	0.000136*	0.000337***
	(0.0000702)	(0.0000715)	(0.0000695)	(0.0000743)	(0.0000838)	(0.0000793)	(0.0000707)	(0.0000709)	(0.0000790)
Observations	1085020	1085020	1085020	1085020	1085020	1085020	1085020	1085020	1085020
Outcome Mean	0.00123	0.00123	0.00123	0.00123	0.00123	0.00123	0.00123	0.00123	0.00123
Assigned Div by Cohort FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Recruit Characteristics	Y	Y	Y	Y	Y	Y	Y	Y	Y
Call Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y
FTO Gender	-	Y	-	-	-	-	-	-	Y
FTO Race	-		Y	-	-	-	-	-	Y
FTO Age	-	-	-	Y	-	-	-	-	Y
FTO Arrest Rate	-	-	-	-	Y	-	-	-	Y
FTO Misd Arrest Rate	-	-	-	-	-	Y	-	-	Y
FTO Response Time Rate	-	-	-	-	-	-	Y	-	Y
FTO Time on Call Rate	-	-	-	-	-	-	-	Y	Y

Standard errors in parentheses
* $p < .1$, ** $p < .05$, *** $p < .01$

Notes: Standard errors are clustered at the recruit officer level. A one standard deviation increase in FTO use of force is a 138 percent increase.

Table 2.7: Partner Controls: The Effect of High Force Field Training Officers on Recruit Use of Force

	(1) Force	(2) Force	(3) Force
FTO Force Rate	0.000170***	0.000165***	0.000110*
	(0.0000532)	(0.0000450)	(0.0000571)
Observations	1083209	1083209	1083209
Outcome Mean	0.00123	0.00123	0.00123
Assigned Div by Cohort FE	Y	Y	Y
Recruit Characteristics	-	Y	Y
Call Controls	-	-	Y
Partner Controls	Y	Y	Y

Standard errors in parentheses
* $p < .1$, ** $p < .05$, *** $p < .01$

Notes: Standard errors are clustered at the recruit officer level. Column 2 adds controls for recruit characteristics (age, gender, race). We add call characteristics fixed effects (number of officers on the scene, beat, type of call—priority-by-type,year-by-month, and day of the week-by night) in column 3. Each column includes controls for partner gender, age, race and force rate.

Table 2.8: The Effect of High Force Field Training Officers on Recruit Arrests

	(1)	(2)	(3)	(4)	(5)	(6)
	Misd. Arrest	Misd. Arrest	Misd. Arrest	Unfiled Arrest	Unfiled Arrest	Unfiled Arrest
Panel A: Lower Quality Arrests						
Force Rate	0.00118**	0.00103**	0.000568*	0.00105**	0.000913***	0.000441**
	(0.000484)	(0.000412)	(0.000290)	(0.000411)	(0.000339)	(0.000204)
Outcome Mean	0.0215	0.0215	0.0215	0.0143	0.0143	0.0143
	(1)	(2)	(3)	(4)	(5)	(6)
	Felony Arrest	Felony Arrest	Felony Arrest	Filed Arrest	Filed Arrest	Filed Arrest
Panel B: Higher Quality Arrests						
Force Rate	0.000210	0.000224	0.0000871	0.000360	0.000365	0.000235
	(0.000136)	(0.000130)	(0.000113)	(0.000267)	(0.000241)	(0.000227)
Outcome Mean	0.00755	0.00755	0.00755	0.0148	0.0148	0.0148
Observations	1085020	1085020	1085020	1085020	1085020	1085020
Assigned Div by Cohort FE	Y	Y	Y	Y	Y	Y
Recruit Characteristics	-	Y	Y	-	Y	Y
Call Controls	-	-	Y	-	-	Y

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

Notes: Standard errors are clustered at the recruit officer level. Columns 2 and 4 adds controls for recruit characteristics (age, gender, race). We add call characteristics fixed effects (number of officers on the scene, beat, type of call—priority-by-type,year-by-month, and day of the week-by night) in columns 3 and 6. Unfiled arrests are arrests that are not filed with the Dallas District Attorney's Office.

Table 2.9: Reporting Concerns: The Effect of High Force Field Training Officers on Recruit Use of Force

	(1) Force	(2) Force	(3) Force	(4) Force	(5) Force	(6) Force	(7) Force	(8) Force	(9) Force
FTO Force Rate	0.000232***	0.000237***	0.000186***	0.000231***	0.000237***	0.000185***	0.000220***	0.000225***	0.000174**
	(0.0000680)	(0.0000648)	(0.0000711)	(0.0000677)	(0.0000646)	(0.0000708)	(0.0000670)	(0.0000636)	(0.0000697)
Observations	1074474	1074474	1074474	1074474	1074474	1074474	1074474	1074474	1074474
Outcome Mean	0.00124	0.00124	0.00124	0.00124	0.00124	0.00124	0.00124	0.00124	0.00124
Div-Cohort FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Recruit Chars	-	Y	Y	-	Y	Y	-	Y	Y
Call Controls	-	-	Y	-	-	Y	-	-	Y
FTO # Characters	Y	Y	Y	-	-	-	-	-	-
FTO # Words	-	-	-	Y	Y	Y	-	-	-
FTO Write Nothing	-	-	-	-	-	-	Y	Y	Y

Standard errors in parentheses
* $p < .1$, ** $p < .05$, *** $p < .01$

Notes: Standard errors are clustered at the recruit officer level. There are 11 FTOs who are not matched to the incident data, so the number of observations in this Table is 1074474 instead of 1085020. A one standard deviation increase in FTO use of force is a 138 percent increase.

APPENDIX A

SUPPLY SIDE RESPONSES IN SCHOOL CHOICE

A.1 Appendix

Table A.1: DiD Results With Shrunken Value-Added Estimates

VARIABLES	(1) Shrunk School VA	(2) Shrunk Math VA	(3) Shrunk Reading VA
$Post_t \cdot HighExp_s$	0.021*** (0.005)	0.026*** (0.007)	0.014*** (0.005)
Observations	15,348	15,348	15,348
R-squared	0.450	0.438	0.442
Baseline Mean	0.0163	0.0274	0.01

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All standard errors are clustered at the school level. Each coefficient is the result of a separate estimation. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as having at least one nearby choice school. Data on enrollment come from the IDOE-CREO database and school value-added is calculated using Equation (1) and shrunk according to a empirical bayes approach [Kane and Staiger, 2008).

Table A.2: DiD Results Varying School VA Estimation

VARIABLES	(1) Baseline	(2) School-Year FE Only	(3) Including Demographics	(4) Including Previous Test Score
$Post_t \cdot HighExp_s$	0.023*** (0.006)	0.024** (0.010)	0.028*** (0.010)	0.023*** (0.006)
Observations	15,348	15,348	15,348	15,348

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All standard errors are clustered at the school level. Each coefficient is the result of a separate estimation. All regressions include school and year fixed effects. Baseline covariates include the share of students that are section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as being in the top tercile of the distribution of number of nearby public schools. Data on enrollment come from the IDOE-CREO database.

Table A.3: DiD Results With Various Definitions of Nearby Choice School

	(1) Within 3 miles	(2) Within 5 miles	(3) Within 8 miles	(4) Within 10 miles	(5) Within 15 miles
Panel A: School Value-Added					
$Post_t \cdot HighExp_s$	0.025***	0.023***	0.013**	0.009	0.004
	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)
Observations	15,348	15,348	15,348	15,348	15,348
R-squared	0.448	0.448	0.447	0.447	0.447
Baseline Mean	0.0197	0.0197	0.0197	0.0197	0.0197
Panel B: School Math Value-Added					
$Post_t \cdot HighExp_s$	0.033***	0.030***	0.016**	0.011	0.004
	(0.007)	(0.007)	(0.007)	(0.008)	(0.009)
Observations	15,348	15,348	15,348	15,348	15,348
R-squared	0.434	0.433	0.432	0.432	0.432
Baseline Mean	0.0255	0.0255	0.0255	0.0255	0.0255
Panel C: School Reading Value-Added					
$Post_t \cdot HighExp_s$	0.016***	0.013***	0.008	0.004	0.001
	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)
Observations	15,348	15,348	15,348	15,348	15,348
R-squared	0.455	0.455	0.455	0.454	0.454
Baseline Mean	0.00390	0.00390	0.00390	0.00390	0.00390

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All standard errors are clustered at the school level. Each coefficient is the result of a separate estimation. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as having at least one nearby choice school within a certain number of miles as indicated in each of the columns. Data on enrollment come from the IDOE-CREO database and school value-added is calculated using Equation (1).

Table A.4: DiD Results on School VA with and without Baseline Covariates

	(1) School Value-Added	(2) School Value-Added	(3) School Math Value-Added	(4) School Math Value-Added	(5) School Reading Value-Added	(6) School Reading Value-Added
$Post_t \cdot HighExp_s$	0.010**	0.0229***	0.016**	0.0302***	-0.002	0.0127***
	(0.005)	(0.0056)	(0.006)	(0.0071)	(0.004)	(0.0048)
Observations	15,348	15,348	15,348	15,348	15,348	15,348
R-squared	0.444	0.4479	0.429	0.4333	0.450	0.4547
School FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Baseline Covariates	No	Yes	No	Yes	No	Yes
Baseline Mean	0.0197	0.0197	0.0255	0.0255	0.00390	0.00390

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All standard errors are clustered at the school level. Each coefficient is the result of a separate estimation. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as having at least one nearby choice school. Odd columns show results when baseline covariates are excluded from the regression. Even columns show the baseline results with the inclusion of baseline covariates. Data on enrollment come from the IDOE-CREO database and school value-added is calculated using Equation (1).

Table A.5: DiD Results on School VA Using a Continuous Measure

	(1) School Value-Added	(2) School Math Value-Added	(3) School Reading Value-Added
$Post_t \cdot NumClose_s$	0.003***	0.004***	0.002***
	(0.001)	(0.001)	(0.001)
Observations	15,348	15,348	15,348
R-squared	0.448	0.433	0.455
Baseline Mean	0.0197	0.0255	0.00390
Avg. Num. Close Schools	4	4	4

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All standard errors are clustered at the school level. Each coefficient is the result of a separate estimation. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. $NumClose_s$ is a continuous measure of the number of choice schools within 5 miles. Data on enrollment come from the IDOE-CREO database and school value-added is calculated using Equation (1).

Table A.6: Student-Level DiD Results of Voucher Program

VARIABLES	(1) Standardized Test Score	(2) Standardized Math Score	(3) Standardized Reading Score
Panel A: High Exposure Public Students vs. Control			
$Post_t \cdot HighExp_s$	0.011**	0.017***	0.003
	(0.004)	(0.006)	(0.004)
Observations	3,753,591	3,753,591	3,753,591
R-squared	0.774	0.705	0.670

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All standard errors are clustered at the school level. Each coefficient is the result of a separate estimation. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as having at least one nearby choice school. Data on enrollment come from the IDOE-CREO database and school value-added is calculated using Equation (1).

Table A.7: DiD Results by Title I Status

VARIABLES	(1) School Value-Added	(2) School Math Value-Added	(3) School Reading Value-Added
Panel A: Had Title I Program			
$Post_t \cdot HighExp_s$	0.018***	0.029***	0.005
	(0.007)	(0.008)	(0.006)
Interaction with Had Title I Program in 2010	0.010	0.002	0.017**
	(0.008)	(0.009)	(0.007)
Observations	15,348	15,348	15,348
R-squared	0.449	0.434	0.457
Baseline Mean	0.0201	0.0258	0.00441
Panel B: Close to Title I Eligibility Threshold			
$Post_t \cdot HighExp_s$	0.023***	0.031***	0.012**
	(0.006)	(0.008)	(0.005)
Interaction with Close to Title I Eligibility Threshold	-0.001	-0.007	0.005
	(0.008)	(0.009)	(0.006)
Observations	15,348	15,348	15,348
R-squared	0.449	0.434	0.457
Baseline Mean	0.0201	0.0258	0.00441

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All standard errors are clustered at the school level. Each coefficient is the result of a separate estimation. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as having at least one nearby choice school. Panel A includes an interaction term that indicates whether a high-exposure public school had a Title I program in 2010. Panel B includes an interaction term that indicates whether a high-exposure public school was within 5 p.p. of the cutoff for Title I eligibility. Data on enrollment come from the IDOE-CREO database and school value-added is calculated using Equation (1).

Table A.8: DiD Results on the Set of Control Schools

VARIABLES	(1) School Value-Added	(2) School Math Value-Added	(3) School Reading Value-Added
$Post_t \cdot HighExp_s$	-0.000	-0.002	0.002
	(0.009)	(0.011)	(0.008)
Observations	6,636	6,636	6,636
R-squared	0.464	0.453	0.450
Baseline Mean	0.0176	0.0261	-0.00236

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All standard errors are clustered at the school level. Each coefficient is the result of a separate estimation. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as having at least one choice school within 8 miles. Data on enrollment come from the IDOE-CREO database.

Table A.9: DiD Results Removing Public Schools With Choice School Within 3-8 Miles

VARIABLES	(1) School Value-Added	(2) School Math Value-Added	(3) School Reading Value-Added
$Post_t \cdot HighExp_s$	0.024***	0.031***	0.015***
	(0.007)	(0.008)	(0.006)
Observations	11,844	11,844	11,844
R-squared	0.441	0.427	0.441
Baseline Mean	0.0143	0.0196	-0.00245

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All standard errors are clustered at the school level. Each coefficient is the result of a separate estimation. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as having at least one choice school within 3 miles. Control public schools are those that do not have a choice school within 8 miles. Data on enrollment come from the IDOE-CREO database. One school is missing data for all years on attendance, so the number of observations is slightly less than other tables.

Table A.10: DiD Results on School VA Dropping Marion County

VARIABLES	(1) School Value-Added	(2) School Math Value-Added	(3) School Reading Value-Added
$Post_t \cdot HighExp_s$	0.019***	0.026***	0.010**
	(0.006)	(0.007)	(0.005)
Observations	13,680	13,680	13,680
R-squared	0.450	0.436	0.455
Baseline Mean	0.0225	0.0272	0.00665

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All standard errors are clustered at the school level. Each coefficient is the result of a separate estimation. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as having at least one nearby choice school. Data on enrollment come from the IDOE-CREO database and school value-added is calculated using Equation (1).

Table A.11: DiD Results Dropping Each County in Indiana

County Dropped	Estimate	Standard Dev.	Observations	County Dropped	Estimate	Standard Dev.	Observations
Adams	0.019***	(0.006)	13,644	Lawrence	0.020***	(0.006)	13,536
Allen	0.020***	(0.006)	12,924	Madison	0.019***	(0.006)	13,572
Bartholomew	0.019***	(0.006)	13,548	Marion	0.019***	(0.006)	13,680
Benton	0.019***	(0.006)	13,668	Marshall	0.020***	(0.006)	13,572
Blackford	0.019***	(0.006)	13,668	Martin	0.019***	(0.006)	13,680
Boone	0.018***	(0.006)	13,536	Miami	0.019***	(0.006)	13,632
Brown	0.019***	(0.006)	13,656	Monroe	0.020***	(0.006)	13,476
Carroll	0.019***	(0.006)	13,668	Montgomery	0.019***	(0.006)	13,584
Cass	0.018***	(0.006)	13,608	Morgan	0.019***	(0.006)	13,500
Clark	0.019***	(0.006)	13,464	Newton	0.018***	(0.006)	13,644
Clay	0.019***	(0.006)	13,596	Noble	0.019***	(0.006)	13,584
Clinton	0.019***	(0.006)	13,620	Ohio	0.019***	(0.006)	13,692
Crawford	0.019***	(0.006)	13,656	Orange	0.019***	(0.006)	13,644
Daviess	0.020***	(0.006)	13,596	Owen	0.019***	(0.006)	13,644
Dearborn	0.019***	(0.006)	13,584	Parke	0.018***	(0.006)	13,620
Decatur	0.019***	(0.006)	13,632	Perry	0.019***	(0.006)	13,656
Dekalb	0.019***	(0.006)	13,584	Pike	0.019***	(0.006)	13,668
Delaware	0.020***	(0.006)	13,452	Porter	0.020***	(0.006)	13,224
Dubois	0.019***	(0.006)	13,548	Posey	0.017***	(0.006)	13,620
Elkhart	0.018***	(0.006)	13,212	Pulaski	0.019***	(0.006)	13,656
Fayette	0.019***	(0.006)	13,620	Putnam	0.019***	(0.006)	13,596
Floyd	0.019***	(0.006)	13,560	Randolph	0.019***	(0.006)	13,584
Fountain	0.019***	(0.006)	13,656	Ripley	0.019***	(0.006)	13,584
Franklin	0.019***	(0.006)	13,656	Rush	0.019***	(0.006)	13,668
Fulton	0.018***	(0.006)	13,644	Scott	0.019***	(0.006)	13,620
Gibson	0.019***	(0.006)	13,596	Shelby	0.019***	(0.006)	13,572
Grant	0.019***	(0.006)	13,548	Spencer	0.019***	(0.006)	13,620
Greene	0.020***	(0.006)	13,596	St. Joseph	0.022***	(0.006)	13,404
Hamilton	0.018***	(0.006)	13,140	Starke	0.019***	(0.006)	13,644
Hancock	0.020***	(0.006)	13,560	Steuben	0.019***	(0.006)	13,608
Harrison	0.019***	(0.006)	13,572	Sullivan	0.019***	(0.006)	13,632
Hendricks	0.018***	(0.006)	13,416	Switzerland	0.019***	(0.006)	13,668
Henry	0.019***	(0.006)	13,548	Tippecanoe	0.018***	(0.006)	13,392
Howard	0.018***	(0.006)	13,524	Tipton	0.019***	(0.006)	13,656
Huntington	0.019***	(0.006)	13,608	Union	0.019***	(0.006)	13,668
Jackson	0.019***	(0.006)	13,560	Vanderburgh	0.017***	(0.006)	13,392
Jasper	0.019***	(0.006)	13,644	Vermillion	0.020***	(0.006)	13,644
Jay	0.020***	(0.006)	13,596	Vigo	0.020***	(0.006)	13,416
Jefferson	0.020***	(0.006)	13,632	Wabash	0.021***	(0.006)	13,608
Jennings	0.019***	(0.006)	13,620	Warren	0.019***	(0.006)	13,656
Johnson	0.017***	(0.006)	13,380	Warrick	0.019***	(0.006)	13,560
Knox	0.017***	(0.006)	13,608	Washington	0.019***	(0.006)	13,644
Kosciusko	0.018***	(0.006)	13,512	Wayne	0.019***	(0.006)	13,536
LaGrange	0.020***	(0.006)	13,680	Wells	0.019***	(0.006)	13,620
LaPorte	0.021***	(0.006)	13,380	White	0.019***	(0.006)	13,608
Lake	0.016***	(0.006)	12,684	Whitley	0.019***	(0.006)	13,608

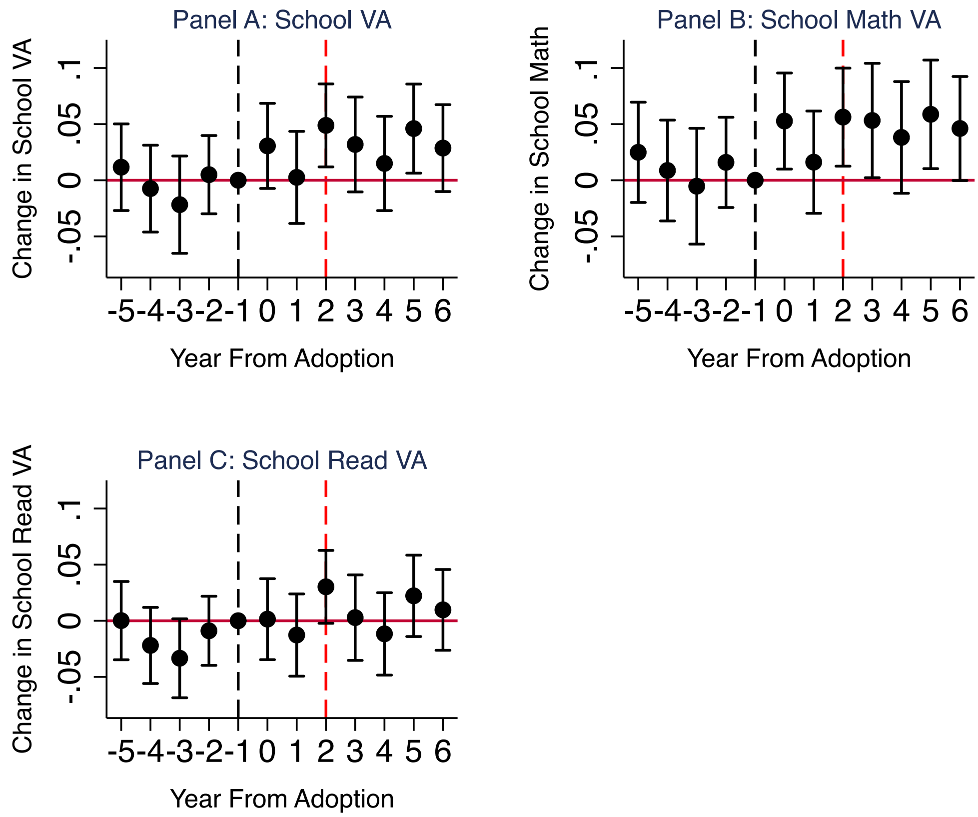
Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All standard errors are clustered at the school level. Each coefficient is the result of a separate estimation. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as having at least one nearby choice school. Data on enrollment come from the IDOE-CREO database and school value-added is calculated using Equation (1).

Table A.12: DiD Results Using Choice Schools - Varying Definition of High Exposure

VARIABLES	(1) School Value-Added	(2) School Math Value-Added	(3) School Reading Value-Added
$Post_t \cdot HighExp_s$	-0.009	-0.019	-0.002
	(0.016)	(0.020)	(0.013)
Observations	2,136	2,136	2,136
Baseline Mean	0.0490	0.0319	0.0725

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. All standard errors are clustered at the school level. Each coefficient is the result of a separate estimation. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as having an above median share of public school students within five miles qualifying for the voucher in 2010. Data on enrollment come from the IDOE-CREO database and school value-added is calculated using Equation (1).

Figure A.1: Event-Study Results of Voucher Policy - High Share of FRPL



This figure presents the event-study estimates from Equation (3). Figure A.1(a) plots the estimates for overall school value-added, Figure A.1(b) plots the estimates for school math value-added and Figure A.1(c) plots the estimates for school reading value-added. Each figure is the result of a separate estimation. 95% confidence intervals are reported. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as having at least one nearby choice school. Data on enrollment come from the IDOE-CREO database and school value-added is calculated using Equation (1).

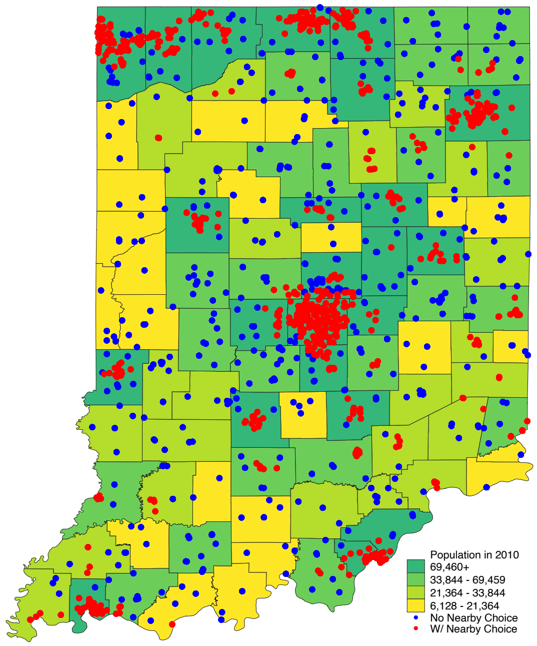
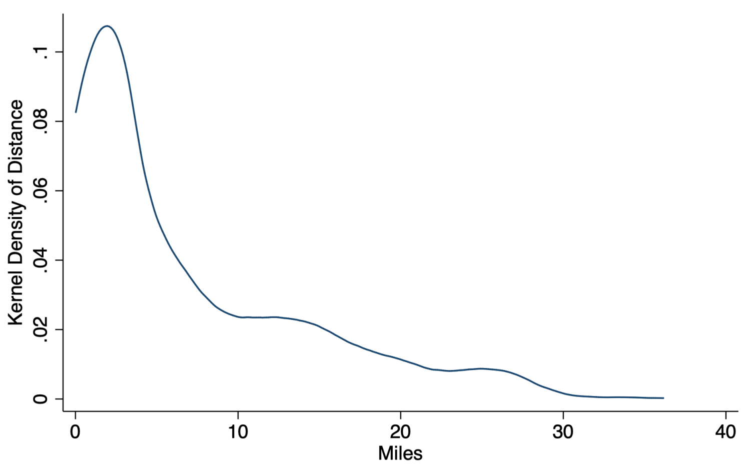


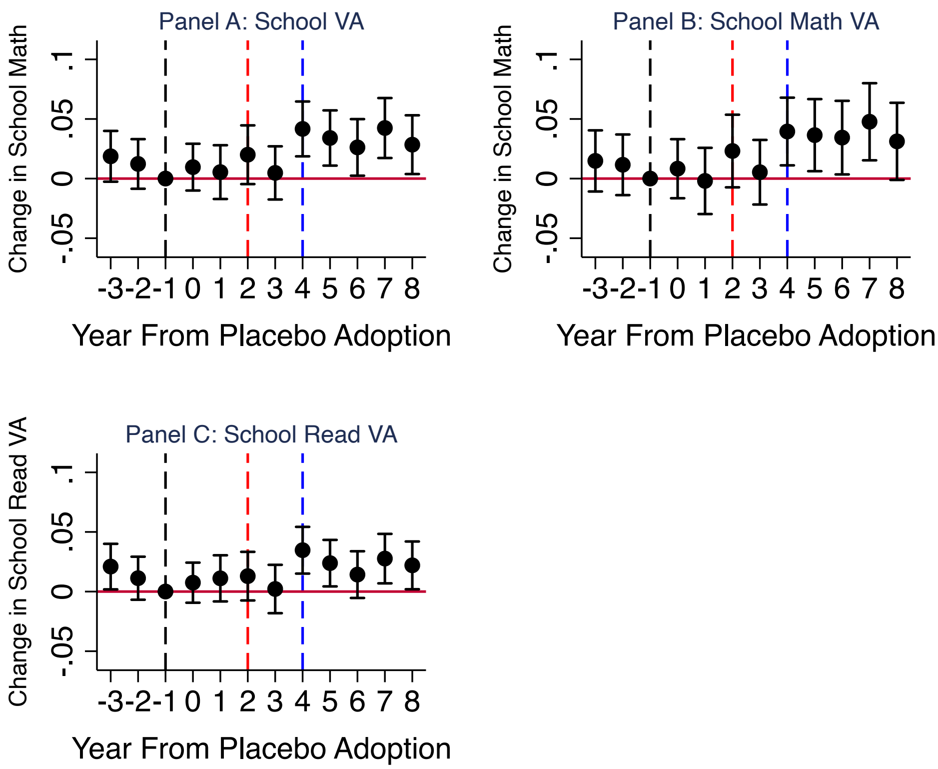
Figure A.2: Locations of High Exposure and Control Public Schools
This figure plots the location of each public school in our sample across Indiana. The red dots indicate the public schools that have a choice school within 5 miles of its location. The blue dots represent the public schools in our control group. The map also shows the population counts for each county in the state in the year 2010. Yellow counties are the least populous, while dark green counties are the most populous. Data on the locations of schools comes from IDOE-CREO database and information on population comes from the U.S. Census Bureau, 2010 Census.



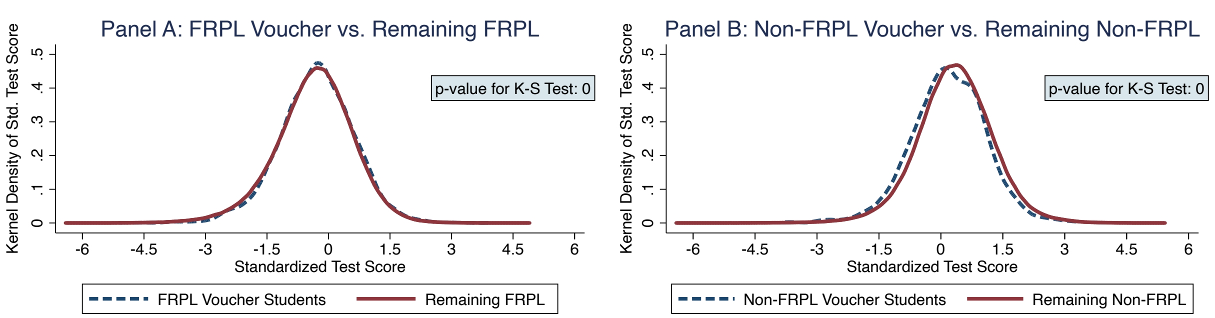
Note: This figure depicts the kernel density plot of the distances between every public school in our sample and the nearest choice school. Distance is caluclated using radial distances between physical addresses. Data on addresses of schools comes from the IDOE-CREO database.

Figure A.3: Kernel Density Plot of Distance to Nearest Choice School

Figure A.4: Placebo Event-Study Results of Voucher Policy

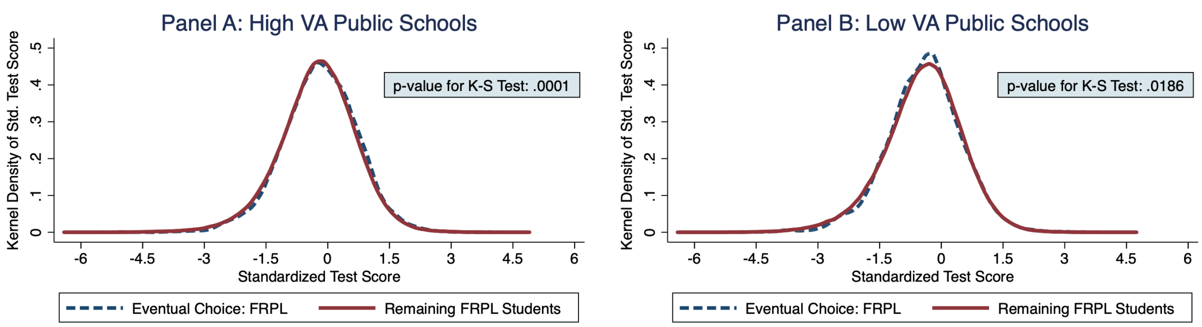


This figure presents the event-study estimates using placebo treatment years. Figure A.4(a) plots the estimates for overall school value-added, Figure A.4(b) plots the estimates for school math value-added and Figure A.4(c) plots the estimates for school reading value-added. Each figure is the result of a separate estimation. 95% confidence intervals are reported. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as having at least one nearby choice school. Data on enrollment come from the IDOE-CREO database and school value-added is calculated using Equation (1).



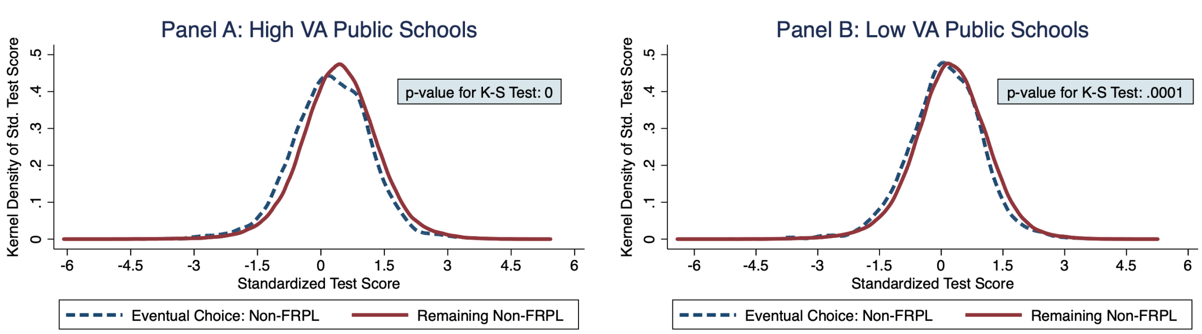
Notes: This figure depicts the kernel density plots of standardized test scores for the students attending public schools in the years before the voucher program. Each panel plots test scores for students who eventually use a voucher and those that remain in public school. Panel A shows the kernel density plots for FRPL eventual voucher students and all FRPL remaining public school students. Panel B shows the kernel density plots for non-FRPL eventual voucher students and non-FRPL students remaining in the public school. Test scores are standardized by year and grade. Data on test scores and enrollment come from the IDOE-CREO database. The p-value for the Kolmogorov-Smirnov equality-of-distributions test is reported for each panel.

Figure A.5: Student Sorting Across FRPL Status



Notes: This figure depicts the kernel density plots of standardized test scores for the students attending public schools in the years before the voucher program. Each panel plots the test scores for FRPL students who use a voucher and those that remain in public school despite qualifying for a 90% voucher. Panel A shows the kernel density plots for students in initially high value-added public schools. Panel B shows the kernel density plots for students in intiatially low value-added public schools. Test scores are standardized by year and grade. Data on test scores and enrollment come from the IDOE-CREO database. The p-value for the Kolmogorov-Smirnov equality-of-distributions test is reported for each panel.

Figure A.6: Kernel Density Plots of Students in Initially High vs Low VA Public Schools: FRPL



Notes: This figure depicts the kernel density plots of standardized test scores for the students attending public schools in the years before the voucher program. Each panel plots the test scores for students who eventually use a voucher and those that remain in public school. Neither group ever qualifies for FRPL. Panel A shows the kernel density plots for students in initially high VA public schools. Panel B shows the kernel density plots for students in intiatially low VA public schools. Test scores are standardized by year and grade. Data on test scores and enrollment come from the IDOE-CREO database. The p-value for the Kolmogorov-Smirnov equality-of-distributions test is reported for each panel.

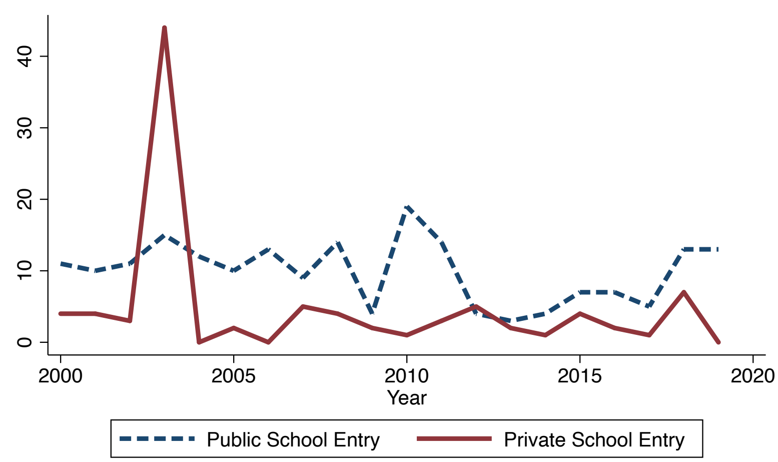
Figure A.7: Kernel Density Plots of Students in Initially High vs Low VA Public Schools: Non-FRPL

A.2 Appendix - Entry and Exit of Schools

Entry and exit into the market is an important supply-side response to consider when evaluating a voucher program. The goal of this appendix section is to illustrate that the Indiana Choice Scholarship Program did not induce significant entry or exit for either public or private schools.

Entry of Public and Private Schools

To understand how entry of schools has changed over the last twenty years, Figure A.8 plots the number of new public and private schools opening in each year from 2000-2019. We only include those schools serving grades third through eighth to match our analysis sample. On average 10 new public schools and 5 new choice schools are opened each year across the state. There is a large spike in the number of private schools in the year 2003. However, there does not seem to significant change in this trend following the adoption of the Indiana Choice Scholarship. We, therefore, conclude that the voucher program did not induce meaningful changes in school entry.

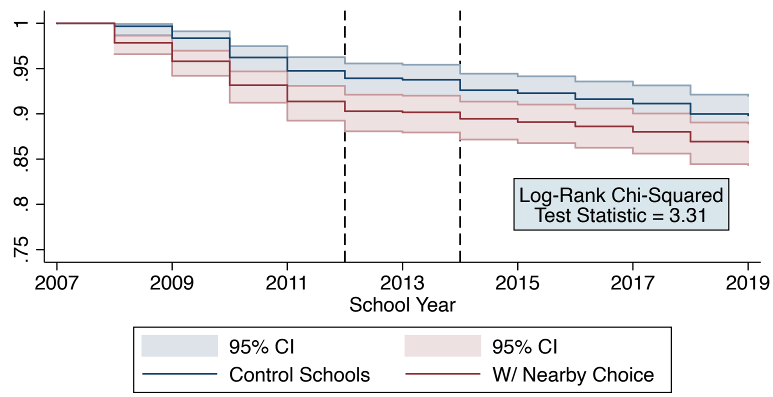


Notes: This figure depicts the number of public and private schools that opened in the state of Indiana in each year from 2000-2019. Data is only shown for schools th cater to grades 3-8. Data on school openings comes from the historical school directories from the IDOE-CREO database.

Figure A.8: Number of Entering Schools by Year

Closures of Traditional Public Schools

We pay particular attention to the closure of traditional public schools because as discussed in Chen and Harris [2021) these events could induce student sorting that biases our results.[1] We investigate this potential mechanism by first examining if public schools located near a choice school saw a differential increase in their likelihood to close. This is done by estimating a Kaplan-Meier survivor function as shown in Figure A.9. It is visually apparent that being located within five miles of an eventual choice school did not increase the likelihood that the public school would close.



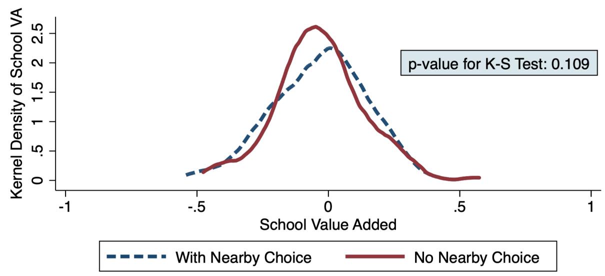
Notes: This figure depicts the Kaplan-Meier survivor function for the set of public schools that existed in Indiana at the start of the sample period. The graphs are separated by whether or not the observation has a near-by choice shcool. A choice school must be within five miles of a public school to be considered near-by. The chi-squared test statistic for the log-rank test of equality is reported. The dashed lines represent the years the voucher program was implemented and expanded.

Figure A.9: Survivor Model for Public School Closures

We next examine whether the quality of the closed public schools differed between those with a nearby choice school and those without one. If the high exposure public schools in our sample receive students from higher quality, closed public schools (when compared to the control group), our estimates may be biased upward. We, therefore, examine whether the distributions of school value-added are equal between these two groups of closed public schools in the years

[1]Chen and Harris [2021) explore this idea in the context of charter school penetration.

before they close. Figure A.10 plots the kernel density functions for the set of public schools that close throughout our sample period. The kernel density functions are estimated separately for the (eventually closed) public schools within five miles of a choice school and those without a nearby choice competitor. It may seem that the school value-added is higher for the set of (eventually closed) public schools within five miles of a choice school, but we fail to reject the null hypothesis of the Kolmogorov-Smirnov equality of distributions test at the 10% level. We take this as suggestive evidence that the quality of the closed public schools did not differ based on the distance to the nearest choice school.

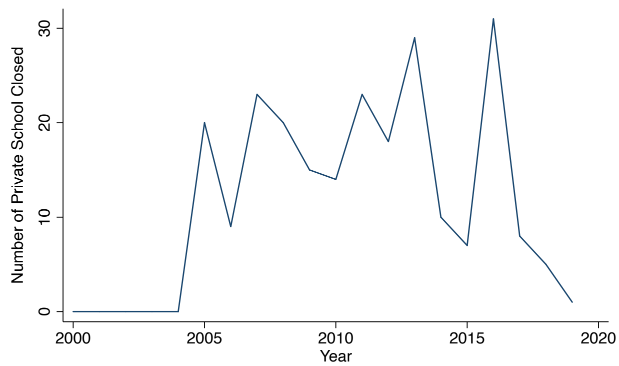


Notes: This figure depicts the kernel density plots of our school VA estimates for the set of public schools that closed during the sample period. The blue line shows the kernel density plot for the schools that have a nearby choice competitor and the red line shows the same for the schools that do not. A near by competitor is defined as having a choice school within 5 miles of the school's location. School VA estimates are calculated using the OLS regression described by Equation (1). Data on test scores and enrollment come from the IDOE-CREO database. The p-value for the Kolmogorov-Smirnov equality-of-distributions test is reported.

Figure A.10: Kernel Density Plot of School VA- Closed Schools

Closures of Private Schools

We also examine the closures of private schools across the state. Figure A.11 plots the number of private schools that close in each year from 2000-2019. We only include schools serving grades third through eighth to match our analysis sample. On average 15 private schools close each year across the state (not including the years that zero schools close). There does not seem to be a systematic change in the trend of private school closures as the program is adopted. We, therefore, conclude that we do not have strong evidence to suggest that the voucher policy induced meaningful changes in the likelihood that a private school would close.



Notes: THis figure depicts the number of private schools that closed in the state of Indiana between 2000 and 2019. Data is only schown for schools that cater grades third through eight. Data on school closures comes from the historical school directories from the IDOE-CREO database.

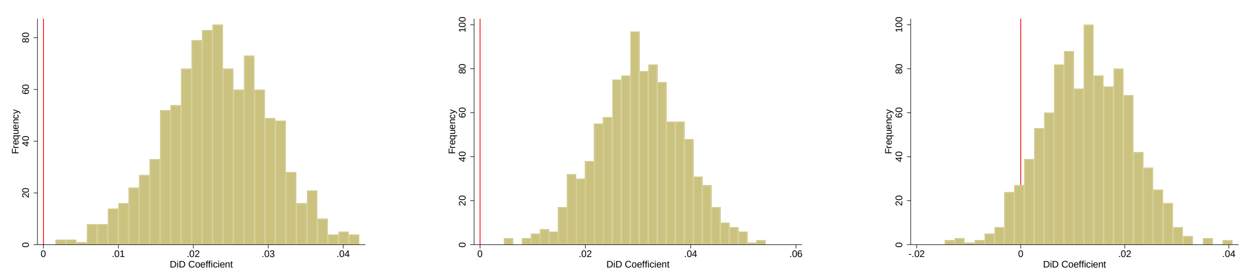
Figure A.11: Number of Private School Closures by Year

A.3 Appendix - Bootstrapping

Deeb [2021) shows that when value-added is the outcome variable of interest in a regression, the regression's robust standard errors used to draw inference are invalid. We, therefore, propose a bootstrapping procedure to correct this issue for our public school analysis.

In each of 1000 iterations, we sample 100 students (with replacement) within each school to be included in the value-added regression described in Equation (1). Therefore, each iteration returns a unique set of school value-added measures that we can use in our difference-in-differences specification. Given this set up, we run Equation (2) on each set of unique school value-added estimates and plot the results on a histogram. We construct new confidence intervals using the standard error of this distribution of difference-in-differences results. We perform this exercise separately for overall school value-added, math school value-added and reading school value-added.

Figure A.12 depicts the results of this exercise. Each panel shows the histogram of difference-in-differences results for each of our outcome variables of interest. We highlight where in the distribution the coefficient equals zero to give a sense of the number of iterations that resulted in the voucher program having zero effect on high-exposure public schools. Across all of the panels, it is evident that a majority (if not all) iterations resulted in positive effect of the program on high-exposure public schools. Table A.13 displays our standard difference-in-differences coefficients along with our newly constructed confidence intervals.



(a) School Overall VA (b) School Math VA (c) School Reading VA

Figure A.12: Bootstrapped DiD Results

Table A.13: DiD Results with Bootstrapped Confidence Intervals

VARIABLES	(1) School Overall Value-Added	(2) School Math Value-Added	(3) School Reading Value-Added
Standard DiD Estimate	0.023	0.030	0.013
Bootstrapped 95% CI	[0.0227, 0.0236]	[0.0298, 0.0308]	[0.0122, 0.0132]

Each coefficient is the result of a separate estimation. All regressions include school and year fixed effects. Baseline covariates include the share of students that are female, white, black, section 504, special education, and receive testing accommodations in the 2006-2007 academic year. High exposure is defined as having at least one nearby choice school. Confidence intervals are calculated according to the procedure describe in A.3.

A.4 Appendix - Structural Model Specification

A.4.A Demand

We will follow the literature on school choice and model a student's schooling decision as a standard discrete choice logit model. All students decide to attend a school j from $j \in \{1, ..., j, ..., S\}$, where S is the number of schools in the market. Students also have restricted choice sets. More specifically, they can only decide between attending a private school or their zoned public school.

Specifically, we model a student i's random utility for attending school j by the following:

$$U_{ij} = \alpha_i P_{ij} + \lambda_i d_{ij} + \gamma_i q_j + \xi_j + \epsilon_{ij} \tag{A.1}$$

where $P_{ij} = P_j - e_i \cdot Voucher_{ij}(I_i, P_j)$ and $e_i = \mathbb{1}\{\text{student } i \text{ is eligible for voucher}\}$, d_{ij} is the distance from student i's home to school j, q_j is the quality of school j (measured as school value-added), ξ_j are unobserved characteristics for school j and ϵ_{ij} is a random preference shock that is distributed as extreme value type 1.

Furthermore, we define mutually exclusive observable groups based on voucher eligibility (income) by which we allow taste parameters to vary.

$$\text{S90} :- \{\text{Set of students eligible for a } 90\% \text{ voucher}\}$$
$$\text{S50} :- \{\text{Set of students eligible for a } 50\% \text{ voucher}\}$$

Then,

$$\alpha_i = \alpha + \alpha_1 \mathbb{1}\{i \in S90 \cup S50\}$$
$$\lambda_i = \lambda_1 \mathbb{1}\{i \in S90\} + \lambda_2 \mathbb{1}\{i \notin S90\}$$
$$\gamma_i = \gamma + \gamma_1 \mathbb{1}\{i \in S90\} + \gamma_2 \mathbb{1}\{i \in S50\} + \rho \nu_i$$

where ν_i is distributed as lognormal(0,1). We use a lognormal as our random coefficient distribution mainly based on findings in previous literature that all households value quality positively. This specification allows preferences for price, quality and distance to differ between high and low-income households. Our random coefficient on quality allows us to fit more flexible substitution patterns for households.

Define $V_{ij} = \alpha_i P_{ij} + \lambda_i d_{ij} + \gamma_i q_j + \xi_j$, then by integrating out the ϵ_{ij} from U_{ij} we get the standard logit CCPs for student i choosing school j given by the following:

$$S_{ij} = \frac{e^{V_{ij}}}{\sum\limits_{k \in C_i} e^{V_{ik}}} \tag{A.2}$$

where C_i is student i's choice set as discussed previously. Note that V_{ij} has to be normalized to 0 for a school in the choice set since there is no explicit outside option.

We obtain market shares by aggregating individual choice probabilities for each school j

$$S_j = \frac{1}{n} \sum_i S_{ij} \tag{A.3}$$

There are two implicit assumptions we make with this choice of modeling. First, all households perfectly observe and make decisions with respect to school quality (school value-added). While it is debated whether this is true for school value-added, we are following the literature by using this measure. Second, we assume capacity constraints for schools are not binding. Although we do not have data on school capacity, conversations with those engaged with schools in Indiana suggest that capacity for choice schools was not an issue during this time.

A.4.B Supply Side

Objective Function of Choice Schools

We assume that private schools are profit maximizers when we model supply. However, there are reasons to believe schools might set tuition and quality under a different objective function. We justify our modeling choice by highlighting the instances in which schools have behaved like profit maximizers and that a large proportion of choice schools' revenue comes from the tuition charged when enrolling students. Below is a quote from an NPR article suggesting that private schools in Indiana were concerned with profits when making decisions for the school [Turner et al., 2017).

> *"We've been seeing some financial troubles here at St. Jude Parish," Runyon said in a formal presentation that was recorded in 2014 and posted on the church's website. The parish was in its third straight year of financial losses. One big reason for the losses: The church was pouring money from its offertory into the school and neglecting repairs to its steeple and cooling system...*
>
> *Not long after, the program was expanded dramatically to include children who had never attended a public school. Suddenly, many St. Jude students qualified. All they had to do was apply. "The effect on that this year," Runyon told parishioners in 2014, "it would have been $118,000 of money we just left there, that the state of Indiana wanted to give me, and we weren't able to take advantage of it."*
>
> *Runyon's presentation — since taken down from the church's website — was a pitch for a new way of distributing financial aid to St. Jude students, one that would maximize the money coming in through vouchers and allow the parish to use more of its offertory elsewhere.*

Optimal Tuition and Quality

Following our model assumptions we show private schools' objective function below:

$$\max_{p_j \geq 0, q_j} (p_j - c(q_j))S_j(\mathbf{p}, \mathbf{q}) \tag{A.4}$$

where p_j is the tuition charged by school s, $c(q_j)$ represents marginal cost which is a function of q_j, the quality of school j, and S_j denote the market share which is dependent on the vector of prices and qualities. It is important to note that schools receive the entire tuition amount even if the household receives a voucher. The difference comes from who pays, not how much is paid.

Differentiating the above equation w.r.t p_j and rearranging gives the following first-order condition:

$$p_j^* = c(q_j) - \frac{S_j(.)}{\frac{\partial S_j(.)}{\partial p_j}} \tag{A.5}$$

The pricing F.O.C. resembles that of the traditional Nash-Bertrand model with differentiated products. Here p_j^* is equal to the marginal costs of an additional student and a markup term. As mentioned previously, the marginal cost of an additional student is an increasing function of quality. The markup term is the standard markup term in Nash-Bertrand pricing models which results from schools having local market power.

If we parameterize $c(q_j) = \theta_0 + \theta_1 q_j$, where $\theta_0, \theta_1 > 0$, then by differentiating variable profits w.r.t q_j and rearranging, we get the following first-order condition for optimal quality:[2]

$$q_j^* = \frac{p_j - \theta_0}{\theta_1} - \frac{S_j(.)}{\frac{\partial S_j(.)}{\partial q_j}} \tag{A.6}$$

The F.O.C. for optimal quality is similar to that for optimal tuition. The first term can be thought of as the level of quality chosen if schools had no market power (quality level that achieves zero profits). The second term is the markdown on quality for schools with local market power. This markdown on quality is similar to that in [Neilson, 2021). It is a measure of local market power for each private school. Quality markdowns are a function of the types of students nearby, how they tradeoff quality, OOP expenses, and distance as well as the vector of school attributes including tuition, quality, and locations.

Since we opt to not model public schools directly, we do not have a direct counterpart to a quality markdown for them. However, it can be shown that the quality markdown (for private schools) is inversely proportional to its quality elasticity of demand. We use this relationship as motivation for estimating quality elasticities of demand to understand public school incentives to improve quality under various voucher programs. We estimate the relationship below (Eq. A.7)

[2]Note that we will not need to assume a functional form for marginal costs for our counterfactuals. We do it for notational simplicity.

for the quality elasticity of demand using our estimated demand parameters. Where q_j and S_j represent school j's quality (value added) and market share respectively.

$$\varepsilon_j = \frac{q_j}{S_j(\mathbf{p}, \mathbf{q})} * \frac{\partial S_j(\mathbf{p}, \mathbf{q})}{\partial q_j} \tag{A.7}$$

We use this model framework to estimate changes in enrollment, quality elasticities (for public schools), and quality markdowns (for private schools) as we vary the voucher program holding all other variables constant (i.e tuition, quality, school locations).

A.5 Appendix - Structural Model Estimation

A.5.A Estimation Overview

We estimate parameters from the demand model outlined in appendix D1 using two steps which we detail below. Recall that we defined the indirect utility that student i receives from attending school j is given by $V_{ij} = \alpha_i P_{ij} + \lambda_i d_{ij} + \gamma_i q_j + \xi_j$. We can rewrite this as $V_{ij} = \delta_j + \mu_{ij}$.

Here μ_{ij} captures all features of utility that vary by individual and school, namely, the non-linear individual heterogeneity terms. The δ_j, mean utility for school j, captures everything that doesn't vary within the school (i.e. mean valuation of quality, tuition, as well as unobserved school characteristics).

In the first step of estimation, we use individual-level choice and demographic data to estimate the non-linear parameters and mean utilities using simulated maximum likelihood.

In our second step, we want to estimate the mean valuation of quality and tuition. To do so, we first need to decompose δ_j into observed and unobserved components.

$$\delta_j = \alpha tuit_j + \gamma qual_j + \xi_j \tag{A.8}$$

where $tuit_j$ and $qual_j$ is school j's tuition and quality, respectively. ξ_j are unobserved (to the econometrician) characteristics. Generally, we are worried about the possible correlation of ξ_j with both tuition and quality. To account for this possibility we need to instrument for both tuition and quality as they are endogenous if schools observe the ξ_j prior to choosing tuition and quality.

We focus on third-grade students in South Bend, Indiana from 2011-2018 to estimate the model. In total, this amounts to analyzing school choices by 21,545 students across the 7 school years ($\sim$ 3100 students per year).

A.5.B Simulation Overview

In addition to simulating draws for our random coefficient in the model, we also have some data limitations. Most importantly, we do not observe latitude/longitude coordinates for students who do not use a voucher.[3] To deal with this complication, we simulate locations for those students using moments from the American Community Survey (ACS) data.

We first find the total number of households who are in each public school's catchment region from the ACS data. We next assign 1) public school students to a catchment region due to zoning of schools, and 2) voucher students can be assigned to catchment regions since we know their locations.

Next, we assume that all households in a catchment region are either a public school, voucher, or non-voucher household. With this assumption, we can calculate the number/proportion of households in each catchment region that do not use vouchers.

Now, for every student for whom we do not observe latitude/longitude coordinates, we can assign them a random point within their probabilistically assigned catchment region. We repeat this process 50 times per individual and compute the likelihood function as normal.

A.5.C Identification

We use cross-sectional variation by school year in out-of-pocket expenses students would pay to attend private schools to identify the marginal utility of out-of-pocket expenses. Out-of-pocket expenses can vary with voucher eligibility status which varies within cohorts of students as well as tuition that a school charges which varies each school year. We are not able to separately identify heterogeneity terms for individuals who receive a 50% voucher vs a 90% voucher due to no variation in out-of-pocket expenses for 90% voucher students. This would lead to collinearity with mean utilities hindering our identification.

We identify distance parameters using the cross-sectional variation of observed student locations across cohorts. We do not separately identify distance parameters for 50% voucher-eligible students and students who are not eligible. We opt for this parameterization because of data limitations which were discussed above. Lastly, we identify parameters on quality using variation in value-added both within and across school years.

[3]Note that this means that for eligible students who do not actually use a voucher we would not observe information on locations.

To identify the mean valuations of quality and tuition we employ 2SLS on Equation A.8 above. The first set of instruments we use are the average distance a student in the market must travel to attend their own school and the average distance to a competitor's school. Since we take locations as exogenous in our model distance traveled should be orthogonal to unobserved product characteristics. The second set of instruments we use is the proportion of 90% eligible students who live within 3 miles of their own school and the average proportion of 90% eligible students who live within 3 miles of their competitors' schools.

A.5.D Results

In Table A.14 below, we show the results from our estimated demand model. We find heterogeneity patterns that are very much in line with previous literature that estimates household demand for schools. First, we find that lower-income households are more sensitive to out-of-pocket expenses than their higher-income counterparts. Similarly, we find that lower-income households are more sensitive to traveling further distances than higher-income households. Finally and perhaps most importantly, we find that lower-income households are less sensitive to quality than higher-income households. In fact, lower-income households are quite inelastic with respect to quality.

This rich preference heterogeneity leads to variation in local market power when it comes to a school's incentive to provide quality. Our results suggest that schools near many lower-income students will likely have much weaker incentives to invest in quality than schools in higher-income neighborhoods.

Table A.14: Demand Estimation Results

	α_1	λ_1	λ_2	γ_1	γ_2	ρ	γ	α
Estimate	-4.31	-0.91	-0.56	-3.48	-1.65	0.74	3.73	-0.14
S.E.	0.04	0.01	0.02	0.23	0.31	0.11	7.32	0.28

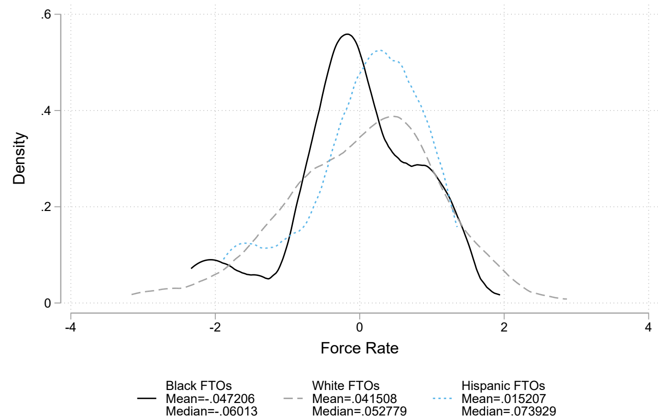
APPENDIX B

THE EFFECT OF FIELD TRAINING OFFICERS ON POLICE USE OF FORCE

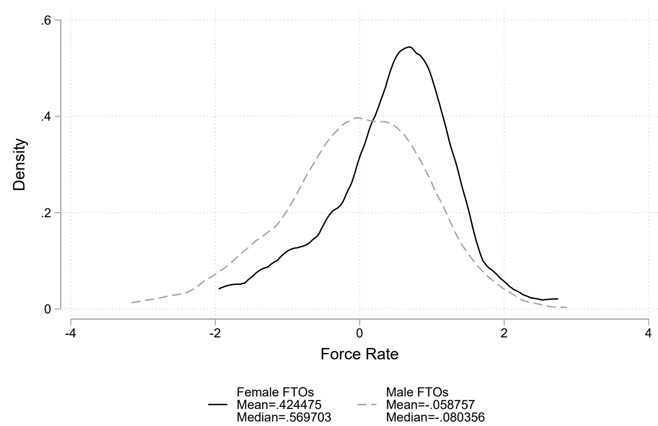
B.1 Appendix

Figure B.1: Density of Field Training Officer Propensity to Use Force by Field Training Officer Characteristics

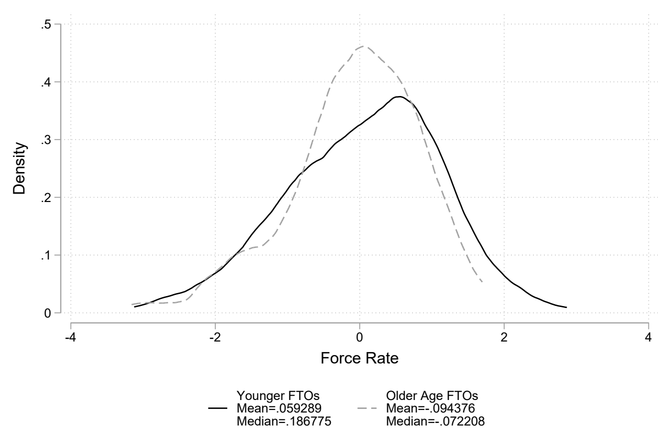
(a) Field Training Officer Race



(b) Field Training Officer Gender

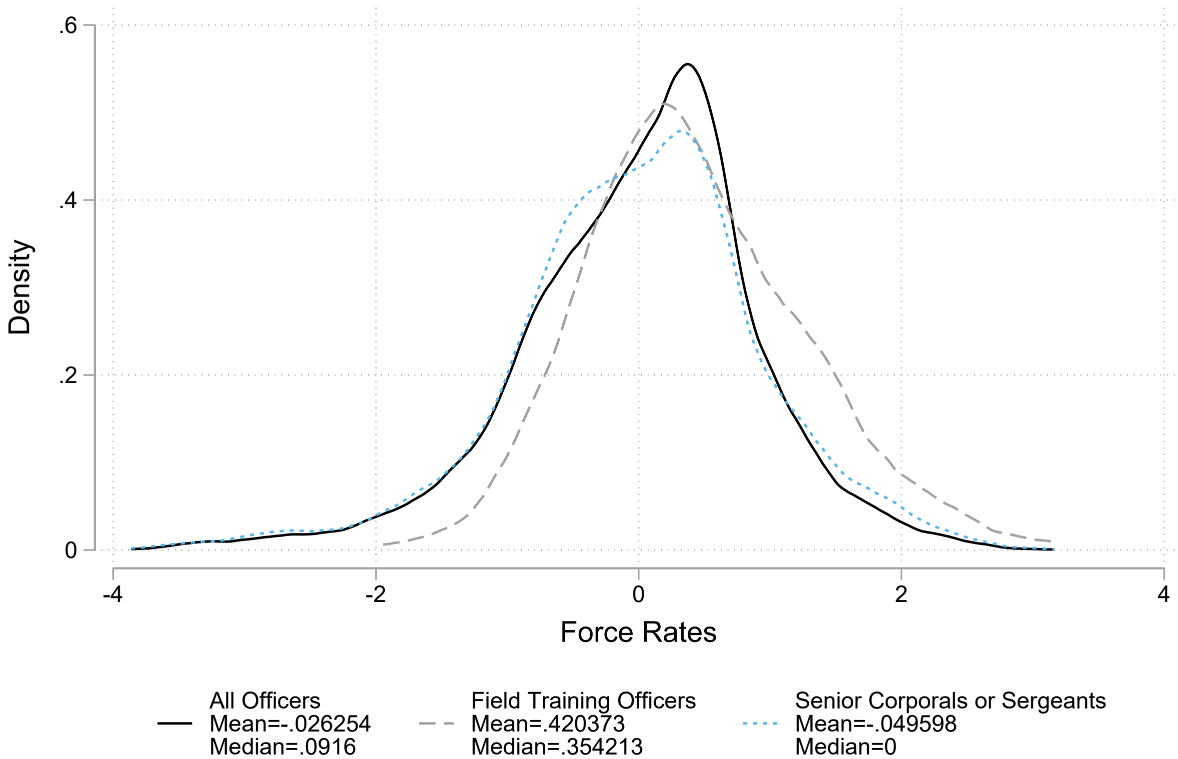


(c) Field Training Officer Age



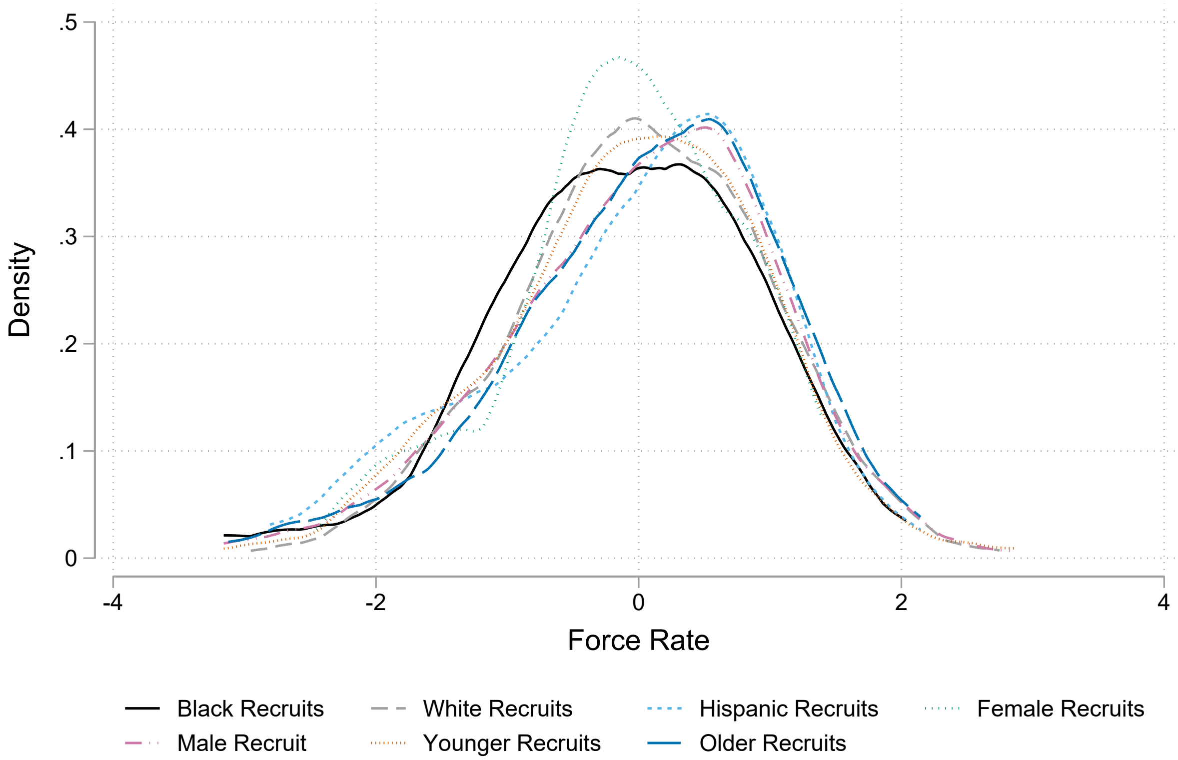
Notes: This figure plots the distribution of field training officer effects by field training officer characteristics.

Figure B.2: Density of Officer Propensity to Use Force for All Officers



Notes: This figure plots the distribution of police officer effects in the full sample of calls. Senior Corporals and Sergeants are the most frequent rank of field training officers.

Figure B.3: Density of Field Training Officer Propensity to Use Force by Recruit Characteristics



Notes: This figure plots the distribution of field training officer effects by recruit characteristics. Older recruits are recruits older than the average age (36 years old).

Table B.1: All Three FTOs: The Effect of High Force Field Training Officers on Recruit Use of Force

	(1) Force	(2) Force	(3) Force
FTO 1 Force Race	0.000199***	0.000207***	0.000154**
	(0.0000673)	(0.0000666)	(0.0000711)
FTO 2 Force Race	0.0000562	0.000112*	0.000101
	(0.0000627)	(0.0000643)	(0.0000666)
FTO 3 Force Race	-0.0000229	0.00000277	0.0000151
	(0.0000673)	(0.0000716)	(0.0000713)
Observations	1085020	1085020	1085020
Outcome Mean	0.00123	0.00123	0.00123
Assigned Div by Cohort FE	Y	Y	Y
Recruit Characteristics	-	Y	Y
Call Controls	-	-	Y

Standard errors in parentheses
* $p < .1$, ** $p < .05$, *** $p < .01$

Notes: Standard errors are clustered recruit level.

Table B.2: Balance Test: Correlation between Recruit and Field Training Officer Characteristics

	(1) Age	(2) Female	(3) White	(4) Black	(5) Hispanic	(6) Hire Date	(7) Force Rate
White	-1.0720	-0.1124	-0.1555	-0.0052	0.0952	281.5651	-0.3147
	(1.8206)	(0.0836)	(0.1025)	(0.0756)	(0.0731)	(912.6772)	(0.2048)
Black	-1.4251	-0.0514	-0.0800	0.0090	0.0061	334.1880	-0.4693**
	(1.9620)	(0.0824)	(0.1159)	(0.0749)	(0.0769)	(955.4714)	(0.2186)
Hispanic	-1.3049	-0.0561	-0.1512	0.0158	0.0730	418.5266	-0.3657
	(1.8905)	(0.0914)	(0.1074)	(0.0810)	(0.0752)	(996.3690)	(0.2424)
Female	-0.1773	0.0293	-0.0923	0.0502	0.0560	289.7770	-0.1121
	(1.1544)	(0.0455)	(0.0597)	(0.0577)	(0.0516)	(415.3940)	(0.1691)
Age	0.0502	-0.0000	0.0027	0.0003	0.0004	10.6934	0.0051
	(0.0953)	(0.0032)	(0.0053)	(0.0042)	(0.0033)	(31.8118)	(0.0112)
Observations	411	411	411	411	411	411	411
Div-x-Cohort FE	X	X	X	X	X	X	X
Outcome Mean	49.45	0.129	0.628	0.170	0.165	15298	-1.51e-09
F-Test P-Value	0.970	0.257	0.0818	0.955	0.331	0.945	0.431

Notes: Standard errors are clustered at the initial assignment-by-cohort year level.

Table B.3: Balance Test: Correlation between Recruit and Field Training Officer Characteristics

	(1) Age	(2) Female	(3) White	(4) Black	(5) Hispanic	(6) Hire Date	(7) Force Rate
White	-1.0720	-0.1124	-0.1555	-0.0052	0.0952	281.5651	-0.3147
	(1.7427)	(0.0845)	(0.0997)	(0.0688)	(0.0705)	(722.8539)	(0.1942)
Black	-1.4251	-0.0514	-0.0800	0.0090	0.0061	334.1880	-0.4693*
	(1.9739)	(0.0939)	(0.1163)	(0.0951)	(0.0801)	(809.8303)	(0.2394)
Hispanic	-1.3049	-0.0561	-0.1512	0.0158	0.0730	418.5266	-0.3657
	(1.8763)	(0.0874)	(0.1094)	(0.0921)	(0.0806)	(760.2191)	(0.2420)
Female	-0.1773	0.0293	-0.0923	0.0502	0.0560	289.7770	-0.1121
	(0.9826)	(0.0433)	(0.0639)	(0.0524)	(0.0538)	(328.7996)	(0.1070)
Age	0.0502	-0.0000	0.0027	0.0003	0.0004	10.6934	0.0051
	(0.0798)	(0.0033)	(0.0049)	(0.0038)	(0.0042)	(29.2205)	(0.0101)
Observations	411	411	411	411	411	411	411
Div-x-Cohort FE	X	X	X	X	X	X	X
Outcome Mean	49.45	0.129	0.628	0.170	0.165	15298	-1.51e-09
F-Test P-Value	0.961	0.515	0.338	0.928	0.389	0.877	0.324

Notes: Standard errors are clustered at the field training officer level.

Table B.4: Robustness Different Force Measures: The Effect of High Force Field Training Officers on Recruit Use of Force

	(1) Force	(2) Force	(3) Force	(4) Force	(5) Force	(6) Force	(7) Force	(8) Force	(9) Force	(10) Force	(11) Force	(12) Force
Theta Shrunk	0.113***	0.116***	0.0846**									
	(0.0375)	(0.0368)	(0.0397)									
Above Avg Theta Shrunk				0.000221	0.000262*	0.000179						
				(0.000137)	(0.000139)	(0.000141)						
IHS Theta Shrunk							0.113***	0.116***	0.0846**			
							(0.0375)	(0.0368)	(0.0397)			
Unshrunken Force Rate										0.0987***	0.0986***	0.0728**
										(0.0322)	(0.0315)	(0.0333)
Observations	1085020	1085020	1085020	1085020	1085020	1085020	1085020	1085020	1085020	1085020	1085020	1085020
Outcome Mean	0.00123	0.00123	0.00123	0.00123	0.00123	0.00123	0.00123	0.00123	0.00123	0.00123	0.00123	0.00123
Assigned Div by Cohort FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Recruit Characteristics	-	Y	Y	-	Y	Y	-	Y	Y	-	Y	Y
Call Controls	-	-	Y	-	-	Y	-	-	Y	-	-	Y

Standard errors in parentheses
* $p < .1$, ** $p < .05$, *** $p < .01$

Table B.5: Robustness Attrition: The Effect of High Force Field Training Officers on Recruit Use of Force

	(1) Force	(2) Force	(3) Force
FTO Force Rate	0.000226***	0.000230***	0.000186**
	(0.0000700)	(0.0000695)	(0.0000732)
Observations	772444	772444	772444
Outcome Mean	0.00128	0.00128	0.00128
Assigned Div by Cohort FE	Y	Y	Y
Recruit Characteristics	-	Y	Y
Call Controls	-	-	Y

Standard errors in parentheses
* $p < .1$, ** $p < .05$, *** $p < .01$

Notes: Standard errors are clustered at the recruit officer level. Column 2 adds controls for recruit characteristics (age, gender, race). We add call characteristics fixed effects (number of officers on the scene, beat, type of call—priority-by-type,year-by-month, and day of the week-by night) in column 3. The sample is only calls within two years of training.

Table B.6: Robustness More Reportable Force: The Effect of High Force Field Training Officers on Recruit Use of Force

	(1) Reportable Force	(2) Reportable Force	(3) Reportable Force
FTO Force Rate	0.0000136**	0.0000115*	0.0000106*
	(0.00000624)	(0.00000595)	(0.00000599)
Observations	1083076	1083076	1082876
Outcome Mean	0.0000507	0.0000507	0.0000508
Assigned Div by Cohort FE	Y	Y	Y
Recruit Characteristics	-	Y	Y
Call Controls	-	-	Y

Standard errors in parentheses
* $p < .1$, ** $p < .05$, *** $p < .01$

B.2 Appendix: Other Field Training Officer Rates

To better understand the mechanism behind our results, and to rule out explanations such as reporting, we calculate field training officer propensity to make arrests, respond to calls in a timely manner, and write up informative reports in a manner similar to our force rate calculations. Namely, we estimate Equation 2.1 using arrest, misdemeanor arrests, felony arrests, filed arrests, unfiled arrests, response time, and time spent on a call as our outcome. We then shrink our FTO-recruit pair estimates of $\lambda_{o(r)}$ according to Equation 2.2.

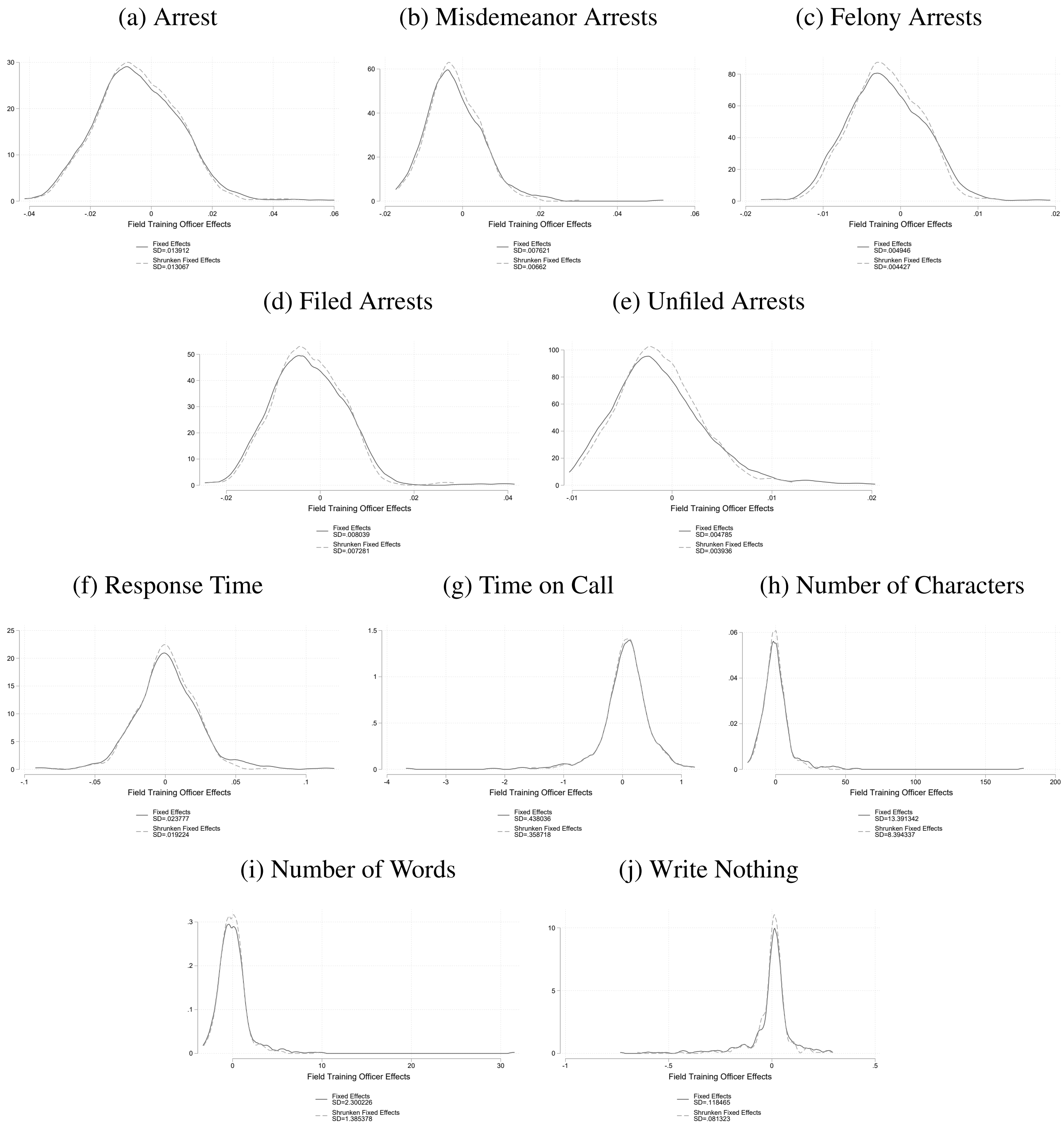
To address whether our results are drive by officer reporting, we also estimate field training officer propensity to write wordy reports. For example, it is reasonable to believe that officers that are more likely to write informative and lengthy reports are also the most likely to report force. Unfortunately we do not have incident reports written by officers for each 911 call. To measure officer wordiness we rely on a separate data set of incident reports. In this data set, we observe 401 of our 411 field training officer-recruit pairs. We attempt to estimate our $\lambda_{o(r)}$'s in a very similar manner, although we do not have exactly the same controls as in our 911 data set. In Equation 2.1, we control for watch instead of night, and we do not control for the number of officer responding to a call. Despite these differences, we believe our analysis is very similar to the what we perform in the 911 sample.

The results of these calculations are show in B.1. Figures B.1a , B.1b, B.1c, B.1d, B.1e show the distribution for our unshrunken and shrunken measures for a field training officer's propensity to make different types of arrests. Both distributions have a longer right tail, indicating that there are some officers with much higher arrest rates than the average field training officer. Further, there is substantial variation in our arrests rates. A one standard deviation increase in officer effects corresponds to a 35% (0.01306/.037)and 31% (0.0066/ 0.0215) increase in arrest or misdemeanor arrest rates.

Figures B.1c and B.1d show results for our measures of time use (measured in hours). A one standard deviation increase in response time is 1 minute (0.02 hours) or 17% (.02/.12 hours) increase. A one standard deviation increase in time on a call is 0.359 hours or a 150% (.359/.24) increase.

Finally, we consider how many words an officer uses when writing up an incident in Figures B.1e, B.1f, and B.1j . Unsurprisingly, the two distributions for number of words and characters look similar. The average number of characters used in an incident report is 43.19976 and there are a few officers that are very wordy. A one standard deviation increase in number of characters used is an increase of 8.39 characters or 19%. The average number of words in a report is 7.31. A one standard deviation increase in wordiness is an increase of 1.39 words or a 20% (1.39/7.31) increase. Finally, we consider an officers propensity to write nothing. On average 10 % of incidents don't have a description. A one standard deviation increase in writing nothing corresponds to a 81% (0.08/0.099) increase in writing nothing. Together these figures show substantial variation in other officer behaviors.

Figure B.1: Other Field Training Officer Rates



Notes: This figure plots the distribution of field training officer effects for arrests, misdemeanor arrests, response time, time on a call and measures of wordiness. Response time is the number of hours between arrival time and assigned time. Time on call is the number of hours between the time an officer was enroute and when the call was cleared.