Essays on the Economics of Education Policy and Regulation

By

Daniel Mangrum

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Economics

June 30th, 2020

Nashville, Tennessee

Approved:

Andrea Moro, Ph.D.

Christopher Carpenter, Ph.D.

Andrew Dustan, Ph.D.

Brent Evans, Ph.D.

Lesley J. Turner, Ph.D.

To my wife, Nina,
my mother, Patty,
and my grandfather, Master Sergeant Thomas Mangrum, Sr.

Acknowledgments

I would like to thank Andrea Moro, Kitt Carpenter, Andrew Dustan, Lesley Turner, Alejandro Molnar, Brent Evans, Andrew Goodman-Bacon, Carly Urban, Michelle Marcus, and Andrew Hussey for their valuable insights and their support of my research. I am also grateful for feedback received from seminar participants at the Southern Economic Association Annual Conference, the Missouri Valley Economic Association Annual Conference, the Association for Education Finance and Policy Annual Conference, and the NBER Summer Institute as well as seminar participants from Vanderbilt University, the University of Memphis, Western Illinois University, the University of Georgia, CNA Corporation, the Bureau of Labor Statistics, the Federal Reserve Bank of New York, the Urban Institute, San Francisco State University, and Middle Tennessee State University. Research in this dissertation was enhanced via financial support from the Kirk Dornbush Summer Research Grant, the Kirk Dornbush Dissertation Fellowship, and from the Washington Center for Equitable Growth. This dissertation would not have been possible if not for the support and encouragement from my family, Nicolas Mäder, Tam Bui, Paul Niekamp, Mathew Knudson, and my other classmates, faculty, and support staff in the Vanderbilt Economics Department. I am especially grateful to the participants in the Vanderbilt Applied Microeconomics Advising Group and the Vanderbilt Economics Job Market Workgroup for their input and advice.

Table of Contents

List of Tables

Table **Page**

List of Figures

Figure **Page**

Chapter 1: Personal Finance Education Mandates and Student Loan Repayment

1.1 Introduction

In the United States, high school students are increasingly tasked with making consequential human capital investment decisions with a rather limited information set. Some of these decisions include whether to attend college, which college to attend, and how to finance postsecondary education. Perhaps as a result of such decision-making under uncertainty, 47% of Americans with student loan debt say they would have accepted fewer federal aid dollars if they could make the choice again. Also, over half say they've had difficulties making monthly payments (Consumer Reports, 2016). These difficulties can be compounded for first generation students who often do not have access to mentors with first-hand experience in postsecondary education. Several recent studies document these frictions and find that these students are less likely to apply to selective universities, are less likely to retake standardized tests, and are more likely to under-invest in postsecondary education (Hoxby and Turner, 2015; Goodman, Gurantz, and Smith, 2018; Avery and Turner, 2012).

In this paper, I study an intervention that might improve outcomes for high school students making human capital investment decisions: mandated personal financial literacy (PFL) coursework. Specifically, I investigate whether PFL education during high school improves postsecondary finance outcomes by providing students with additional information at the critical time financial aid decisions are made. Between 1993 and 2014, 23 states adjusted high school graduation requirements to include topics covering personal financial literacy (Stoddard and Urban, 2019). The main objective for the focus on PFL is to increase the overall financial literacy of the state population, but many of the state standards are beginning to include topics discussing postsecondary education and career research. Since enrollment in these courses often coincides with the timing of the federal financial aid application process, requiring personal finance education in high school can operate as a just-in-time information intervention to improve postsecondary finance outcomes.

I estimate how personal finance education mandates affect federal student loan repayment by exploiting plausibly exogenous variation in university-level exposure to state mandates. When states adjust high school graduation requirements to include PFL topics, universities become increasingly populated by students who were exposed to this course content during high school. Universities are differently affected by changes to state standards because their student bodies have different shares of incoming students

from adopting states. I use this exogenous variation in exposure to state graduation requirements to identify the causal effect of required PFL education in high school.

I find that increased exposure to state adopted PFL standards improves university-level student loan repayment. The effect is largest for students at public universities and especially for first generation students and for students from households earning less than $30,000 per year. The estimates suggest a 5% increase in the probability a low income or first generation student is able to pay down some of their balance within a year of entering repayment. I conduct a counter-factual exercise using these estimates which suggests that mandating PFL standards for all high school students between the 2001 and 2008 graduating classes would have resulted in around 9,000 additional students successfully repaying student loans each year.

I also present evidence in support of the identifying assumptions and I conduct a number of robustness checks. Using a flexible event study specification, I show that universities that were more exposed to PFL mandates were not trending differently than universities less exposed to PFL mandates. I confirm results from the literature that the adopted PFL mandates did not significantly shift students to select different colleges (Stoddard and Urban, 2019). Additionally, I estimate an alternative specification which holds fixed the share of students from each feeder state and relies only on the state adoption of mandates over time. The results from this specification are similar to the baseline specification. Taken together, I conclude that the findings are not driven by a compositional change in university cohorts, but via micro-level improvements in student loan repayment.

I explore several mechanisms by which mandated PFL education may improve federal student loan repayment. First, I test whether improvements in student loan repayment are due to decreases in average loan balances. I find that only high income students change borrowing behavior resulting in roughly 8% lower balances upon entering repayment. Point estimates suggest small declines (less than 3%) for first generation and middle income students but the estimates are not precise. Next, I test whether students bound by mandates are better able to correctly answer financial literacy questions. Across three different surveys, I find no evidence of improvements in financial literacy at the time of survey for those that were bound by the state mandates. I also find no evidence that students bound by personal finance mandates were any more likely to attend college or earn a degree. However, I do find that affected students were more likely to correctly answer questions pertaining to federal student loan regulations. The results suggest that, rather than reductions in borrowing or improvements in financial literacy, the personal finance education mandates studied in this paper may act as a just-in-time information

intervention for students making postsecondary decisions.

The remainder of the paper is organized as follows: Section 2.3 reviews background details of the federal financial aid system and summarizes the previous studies discussing personal finance education. Section 1.3 discusses the potential mechanisms by which personal finance education in high school can influence student loan repayment after college. Section 2.4 details the data used in the analysis. Section 2.5 discusses the empirical strategy and assumptions necessary for identification and inference. Section 2.6 presents the empirical results, tests the identification assumptions, and presents evidence for the tested mechanisms. Section 2.7 concludes the paper. I also present results from various robustness checks and the estimation of alternative specifications in Section 1.8.

1.2 Background

In order to qualify for federal student loans, students must complete a Free Application for Federal Student Aid (FAFSA) which collects details about students and their families including income and asset information. While federal subsidized loans are means-tested based on information from the FAFSA, unsubsidized loans are available to any student.[1] Students also face limits on federal borrowing based on the loan type, year of schooling, university Cost of Attendance (COA), and other financial aid received.

A wealth of evidence from the literature has largely concluded that access to financial aid increases access to higher education for low income students (Dynarski, 2003) and more recent evidence suggests increased student loan borrowing causes higher grades and more completed credits for community college students (Marx and Turner, 2019b). Despite the benefits, studies have been critical of the burdensome bureaucracy and complicated process for applying for and receiving financial aid (Dynarski and Scott-Clayton, 2006; Novak and McKinney, 2011; Bettinger, Long, Oreopoulos, and Sanbonmatsu, 2012; Dynarski and Scott-Clayton, 2013; Scott-Clayton, 2015).[2] Critics often argue that the complicated application process and the multitude of choice options tend to reduce the receipt of aid for students that would otherwise be eligible and encourages students to opt into the default option (Kofoed, 2017; Marx and Turner, 2019a). Even for students that successfully navigate the application stage, complexities surrounding the number and type of repayment plans can lead to issues during repayment (Cox, Kreisman, and Dynarski, 2018; Abraham, Filiz-Ozbay, Ozbay, and Turner, 2018). The evidence presented in this paper is

[1] Subsidized loans are loans in which interest does not accrue while the student is in school while unsubsidized loans begin accruing interest after dispersement.

[2] Castleman, Schwartz, and Baum (2015) summarizes several studies that test interventions designed to improve the decision making process in investing in higher education.

consistent with this literature that finds improvements in outcomes through reducing the barriers to federal financial aid access.

A few recent studies find that interventions that provide students with more information about college applications and federal financial aid improve outcomes for students from disadvantaged backgrounds. Bettinger, Long, Oreopoulos, and Sanbonmatsu (2012) show that providing low income families with assistance completing the FAFSA can dramatically increase the probability of applying for federal aid and increase college enrollment and persistence. Barr, Bird, and Castleman (2016) find that providing information to community college students about federal student loan options can shift borrowers away from higher cost financing which is largely driven by students with lower levels of financial literacy and higher debt balances. Gurantz, Pender, Mabel, Larson, and Bettinger (2019) find that virtual college counseling that targets low and middle income students increased the probability students chose to attend a college with a high graduation rate. Additionally, Castleman and Goodman (2018) find that college counseling can increase low income student enrollment and persistence in less expensive four-year public universities with higher graduation rates. Bettinger and Evans (2019) also find that peer advising from recent college graduates to high school students can increase enrollment in two-year colleges for low income and Hispanic students without reducing four-year enrollment.

One large scale intervention that might improve postsecondary outcomes for disadvantaged students is the addition of personal financial literacy (PFL) education during high school. A few recent articles study the effect of PFL directly, the first being Brown, Grigsby, van der Klaauw, Wen, and Zafar (2016). They investigate how changes in economics, mathematics, and personal finance requirements affect financial outcomes for young people. The results confirm findings in the previous literature that increasing math requirements increases asset levels and incomes for young adults (Goodman, 2019). Additionally, mandated PFL coursework is shown to reduce the amount of delinquent debt held by young adults. They also find the effect of course mandates on credit health grows as mandates mature. This suggests either implementation lags on the part of schools or improvements in teaching efficiency over time. Harvey (2019) and Urban, Schmeiser, Collins, and Brown (2018) also investigate how mandated PFL coursework affects financial outcomes for young people. Harvey (2019) finds that young people bound by mandates are less likely to use alternative financial services which typically carry very high interest rates with high rates of delinquency. Urban, Schmeiser, Collins, and Brown (2018) compare credit report data across mandated and non-mandated young people and find that those who were bound by personal finance education mandates during high school have fewer delinquent credit accounts and higher credit scores.

This paper is most closely related to Stoddard and Urban (2019) which studies how PFL mandates affect the receipt of federal student aid for first year college students. Using a difference-in-differences design with several waves of the National Postsecondary Study of Student Aid (NPSAS), they find that mandated college freshmen are more likely to complete the FAFSA, more likely to borrow from federal sources, more likely to receive grants or scholarships, borrow fewer private loans, and are less likely to carry a credit card balance. They also find that the impact on extensive borrowing of federal loan dollars and the lower likelihood of credit card borrowing is larger for low income students and these students are also less likely to work while enrolled in college.

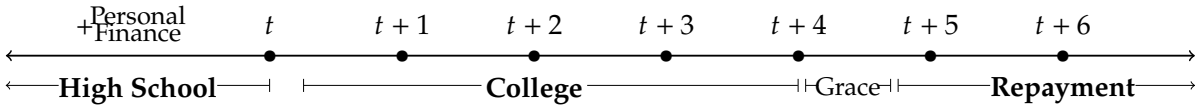
This paper extends this literature in four distinct dimensions. First, to my knowledge, this is the first paper to estimate the impact of PFL mandates on student loan repayment. To this end, I employ two measures of federal student loan repayment progress that vary in sensitivity. The first measure, loan default, is a more adverse and relatively rare outcome that is difficult to affect. The second outcome, the repayment rate, measures whether loan principals are declining and is thus a more sensitive measure of repayment progress. Second, I test whether PFL mandates affect the level of student debt upon leaving college.[3] Third, I investigate the source of these changes by estimating the effect of PFL mandates on financial literacy and on knowledge of the federal student loan system. Lastly, I employ a novel identification strategy to overcome the lack of quality micro-level data that instead relies on university-level benchmarks to proxy borrower level changes in student loan outcomes. I show that, under the necessary assumptions, this identification strategy consistently estimates the micro-level effect of PFL mandates on borrower outcomes.

1.3 Potential Mechanisms

To better understand how mandated personal finance education can influence student loan repayment after college, consider the case of two otherwise identical high school students in which one student (the treated student) is required to be exposed to personal finance education during high school as in Figure 1.1 and the second student (the untreated student) is not. There are many avenues for the intervention to affect repayment outcomes since exposure to personal finance education occurs during high school (t) and student loan repayment does not begin until after college ($t + 5$). The three categories of course content most likely to directly or indirectly affect student loan repayment are topics on financial literacy, financial aid, and career research.

[3] Brown, Grigsby, van der Klaauw, Wen, and Zafar (2016) finds larger student loan balances at age 27 which is a combination of original principal and pace of repayment rather than total loan debt upon entering repayment.

Figure 1.1: Example Timeline for Personal Financial Literacy Education Intervention through to Student Loan Repayment



First, since the goal of most PFL mandates is to improve financial literacy, the standards typically focus on topics such as interest, inflation, risk tolerance, insurance, budgeting, and investing.[4] As a result, students bound by PFL mandates may be better at managing money, be more likely to understand the risks of borrowing, and be more likely to make on time loan payments. The state standards for many courses require students to "examine ways to avoid and eliminate credit card debt" (Texas[5]), "design a financial plan for earning, spending, saving, and investing" (Missouri[6]), and "evaluate the different aspects of personal finance including careers, savings and investing tools, and different forms of income generation" (Michigan[7]). However, most studies that evaluate small scale financial literacy interventions find little to no improvement in financial literacy and most improvements depreciate quickly (Huston, 2010; Fernandes, Lynch Jr, and Netemeyer, 2014). On the other hand, some evidence shows that financial literacy interventions during high school can improve more objective measures of financial health such as credit scores (Brown, Grigsby, van der Klaauw, Wen, and Zafar, 2016; Urban, Schmeiser, Collins, and Brown, 2018).

Second, the required coursework may help students navigate the federal financial aid system. In many states, the PFL standards require students to research various ways of funding postsecondary education. For example, students in Oregon must research the costs and benefits of using loans to finance higher education[8] and students in Texas should "research and evaluate various scholarship opportunities."[9] Some states take this a step further and require students to practice applying for federal financial aid and research the differences in various aid types. Tennessee's state standards require students

[4]Figure 1.13 shows a word cloud of the text of all PFL state standards.

[5]https://web.archive.org/web/20111107152521/http://ritter.tea.state.tx.us/rules/tac/chapter118/ch118a.html

[6]https://dese.mo.gov/sites/default/files/personal_finance_competencies.pdf

[7]https://www.michigan.gov/documents/mde/SS_COMBINED_August_2015_496557_7.pdf

[8]https://www.ode.state.or.us/teachlearn/subjects/socialscience/standards/oregon-social-sciences-academic-content-standards.pdf

[9]https://web.archive.org/web/20111107152521/http://ritter.tea.state.tx.us/rules/tac/chapter118/ch118a.html

to research both positive and negative aspects of borrowing federal student loans[10] while Utah students must "utilize the FAFSA4caster to explore the FAFSA process."[11] Previous research has shown that simplifications in the federal financial aid system often lead to improvements in outcomes for vulnerable groups (Dynarski and Scott-Clayton, 2006; Novak and McKinney, 2011; Bettinger, Long, Oreopoulos, and Sanbonmatsu, 2012; Dynarski and Scott-Clayton, 2013; Scott-Clayton, 2015). Therefore, requiring students to become familiar with the aid process might improve outcomes for low income and first generation students.

Third, many PFL course requirements direct students to research various careers, colleges, and majors. These exercises might alter the trajectory of the student in a number of dimensions. Many of the state standards for the required personal finance education directly address investments in human capital. Students bound by various state mandates are required to "explore potential careers and the steps needed to achieve them" (Arkansas[12]) or to "identify a career goal and develop a plan and timetable for achieving it, including educational/training requirements, costs, and possible debt" (New Jersey[13]). These activities during the personal finance coursework might cause students to be more aware of various career paths and education requirements which might better prepare these students for success in college.

In this paper, I directly test a number of hypotheses that might provide supporting evidence for these mechanisms. I first test whether student loan balances upon entering repayment change as a result of PFL mandates. Next, I test whether PFL mandates improve literacy. Specifically, I test whether students bound by PFL mandates are better able to answer financial literacy questions and questions about federal student loans. As a test for one identification assumption, I investigate whether the adoption of PFL state standards alter students' college choice. Lastly, I test whether students bound by PFL mandates have a higher likelihood of attending college or earning a degree.

1.4 Data

Since micro-level data on repayment outcomes for sequential cohorts are not available to researchers, I instead rely on university-level outcomes. The federal student loan repayment outcomes used in this paper come from the College Scorecard database. The

[10] https://www.tn.gov/content/dam/tn/education/ccte/cte/cte_std_personal_finance.pdf

[11] https://www.schools.utah.gov/file/6348311c-77c7-4fbd-87e7-ba3484bddb6e

[12] http://www.arkansased.gov/public /userfiles/ Learning_Services/Curriculum%20and%20Instruction/Frameworks/Personal_Finance/Economics-aligned-to-PF-Standards.pdf

[13] https://www.state.nj.us/education/cccs/2014/career/91.pdf

College Scorecard was developed during the Obama Administration and debuted in 2015 as a website tool to provide more information to potential college students. The Department of Education provides the underlying university-level data dating back to the 1996-1997 academic year and updates the data annually. The data are sourced via self-reports by universities, from various federal data sources, and from administrative data on students receiving financial aid. The data used in this paper are largely computed using the administrative National Student Loan Data System (NSLDS) which contains records on the universe of federal aid recipients.

I restrict the sample to four-year baccalaureate universities since four-year universities are largely populated by first-time degree seeking recent high school graduates.[14] I remove universities that aggregate repayment outcomes across multiple branch campuses and universities that do not receive federal financial aid.[15] In order to construct a balanced panel, I remove universities that either enter or exit the sample during the sample window. This can occur due to a university opening or closing or a university opting into or losing access to federal aid.[16] The resulting sample contains 1,386 universities across 50 states and the District of Columbia of which 450 are public universities and 936 are private universities.[17]

The two main outcomes from the College Scorecard are the two-year cohort default rate and the one-year repayment rate. After leaving college, federal student loan borrowers are granted a six month grace period before they must begin making monthly payments. One year after entering repayment, borrowers fit into one of four mutually exclusive bins as depicted in Figure 1.2.[18] If the student is making payments on her loans and the balance is declining, this student fits in bin A. If the student is making payments toward her loan but the payment is not sufficient to cover accruing interest (i.e. negative amortization),

[14] In 2016, 45% of recent high school graduates enrolled in 4-year colleges while 23.7% enrolled in 2-year schools. https://nces.ed.gov/fastfacts/display.asp?id=372

[15] This restriction is necessary due to the nature of the identification strategy discussed in the next section. When a university system aggregates outcome measures across multiple branches, the identifying variation on the right-hand-side of the estimating equation is aggregated at a smaller granularity than the outcome measure on the left-hand-side. The reasoning behind the varying level of aggregation is discussed in footnote 17 of *Using Federal Data to Measure and Improve the Performance of U.S. Institutions of Higher Education* found at https://collegescorecard.ed.gov/assets/UsingFederalDataToMeasureAndImprovePerformance.pdf

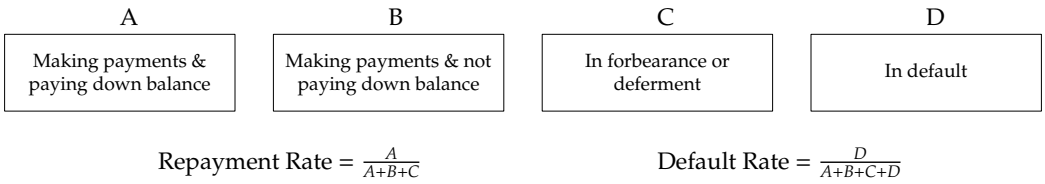
[16] This restriction helps to alleviate the concern of selection into or out of the sample. Universities may lose access to federal aid as a result of poor student loan repayment or choose to begin accepting federal aid as a result of unobservables that change over time. Looney, Yannelis et al. (2019) notes that the majority of the variation in cohort defaults over time stem from entry into and out of the student loan market.

[17] Table 1.12 details the change in sample size as a result of each of the restrictions. Figure 1.12 plots the locations of the universities in the sample.

[18] The College Scorecard database also includes the repayment rate for 3, 5, and 7 years, however the data begins for all variables for FY 2006 and thus these variables have very small windows of data availability.

the student fits in bin B.[19] If the student has been granted forbearance or deferment of payments (and thus no payments are required) and the balance is not declining, the student fits in bin C. If the student has not made any payments for 270 days, the student enters default and fits in bin D.[20]

Figure 1.2: Repayment Status Bins for Repayment Cohort



The two-year **cohort default rate** is calculated by dividing the total number of students in default (bin D) at the end of the two year window by the total number of students in the repayment cohort (the sum of bins A, B, C, and D). The one-year **repayment rate** is calculated by dividing the number of students who have paid down at least one dollar of their original principal (bin A) one year after entering repayment by the total repayment cohort excluding those in default (the sum of bins A, B, and C). Those students in bin B (making payments but facing negative amortization) and in bin C (not required to make payments and facing negative amortization) both count against the repayment rate *but are not in default*. This makes the repayment rate a more sensitive measure of loan repayment health that does not require default but factors in repayment progress.

The College Scorecard reports the default rate and the repayment rate for the full repayment cohort, but the repayment rate is also reported for various subsamples of the student body. Of particular interest in this paper, these subsamples include **first generation students** (students whose parents did not have a college degree upon college entry), **low income students** (students with household income less that $30,000 upon college entry), **middle income students** (students with household income between $30,000 and $75,000 upon college entry), and **high income students** (students with household income above $75,000 upon college entry).[21] Table 1.1 reports summary statistics for the

[19] Due to income-driven repayment plans, it is possible that the monthly minimum payment is not enough to cover the interest accruing each month. In this case, it is very unlikely the borrower will have a declining balance without paying more than the monthly minimum payment.

[20] Default for students loans is atypical compared to other consumer debt. Upon default, there is no repossession of assets since the loans are unsecured. Rather, the federal government levies fines and allows loan services to garnish wages and tax refunds to collect outstanding debts.

[21] The College Scorecard uses nominal dollars to determine these bins. As a result, students with similar household incomes in real terms might be shifted into higher income bins over time due to inflation.

sample of universities for the main outcome variables. I present the means and standard deviations weighted by the number of borrowers used in constructing each university outcome. The weighted moments are presented to be representative of the population of student loan borrowers.

Table 1.1: Descriptive Statistics

	All Universities (n=1,386)		**Public Universities** (n= 450)		**Private Universities** (n= 936)	
Outcome Variable	Mean	SD	Mean	SD	Mean	SD
Default Rate	0.042	(0.032)	0.046	(0.031)	0.036	(0.034)
Repayment Rate						
Overall	0.594	(0.170)	0.593	(0.149)	0.595	(0.199)
First Generation	0.536	(0.165)	0.551	(0.147)	0.512	(0.188)
Low Income (<$30k)	0.458	(0.170)	0.478	(0.151)	0.423	(0.194)
Middle Income ($30k to $75k)	0.631	(0.148)	0.630	(0.136)	0.632	(0.166)
High Income ($75k+)	0.749	(0.114)	0.730	(0.107)	0.777	(0.117)

Means and standard deviations for the main outcome variables are presented above for the full sample and separately by institution control. Moments are weighted by the number of borrowers used to compute each outcome in order to be representative of the population of student borrowers. Default rate is the two year cohort default rate from FY1995 to FY2013. The one year repayment rate is reported for the full repayment cohort and separately for first generation students (students whose parents did not have a college degree) and for students by household income bins.

The repayment rate is first reported in the College Scorecard beginning with the 2007-2008 academic year which includes students entering repayment in the 2006 fiscal year.[22] The most recent data in the Scorecard covers students that entered repayment in the 2013 fiscal year.[23] Beginning in FY2009, the Department of Education began grading universities on the three-year cohort default rate instead of the previous two-year cohort default rate. This change was a concerted effort to hold universities accountable for borrowers beyond two years after entering repayment. The College Scorecard continued to include the two-year cohort default rate through FY2011 but deferred to only posting the three year cohort default rate in subsequent years. Due to the change in the cohort

[22]The College Scorecard reports the one-year repayment rate as a two year rolling average in order to reduce variability. Although this is not ideal for identification, I match the repayment rate outcome using the first year a repayment cohort is reported in the data to match incoming college cohorts to repayment cohorts. Any bias from this rolling average will work *against* detecting an effect of mandates on repayment since it will include one untreated cohort and the first treated cohort.

[23]For repayment cohort counts smaller than 30 students, the data is suppressed and thus these small cells are omitted from the analysis.

default metric, I use the two-year cohort default rate in the available years between FY1995 and FY2011.

Table 1.2: Implementation of Personal Financial Literacy (PFL) Mandates Since 1990

State	Coursework	First Graduating Cohort Bound
New Hampshire	Incorporated (Economics)	1993
New York	Incorporated (Economics)	1996
Michigan	Incorporated (Career Skills)	1998
Wyoming	Incorporated (Social Studies)	2002
Arkansas	Incorporated (Economics)	2005
Arizona	Incorporated (Economics)	2005
Louisiana	Incorporated (Free Enterprise)	2005
South Dakota	0.5 Credit (Economics or Personal Finance)	2006
North Carolina	Incorporated (Economics)	2007
Georgia	Incorporated (Economics)	2007
Idaho	Incorporated (Economics)	2007
Texas	Incorporated (Economics)	2007
Utah	0.5 Credit	2008
Colorado	Incorporated (Economics, Math)	2009
South Carolina	Incorporated (Math, ELA, Social Studies)	2009
Missouri	0.5 Credit	2010
Iowa	Incorporated (21st Century Skills)	2011
Tennessee	0.5 Credit	2011
New Jersey	2.5 Credits (Economics or Personal Finance)	2011
Kansas	Incorporated (Economics)	2012
Oregon	Incorporated (Social Studies)	2013
Virginia	0.5 Credit	2014
Florida	Incorporated (Economics)	2014

PFL mandate data are from Stoddard and Urban (2019). States marked Incorporated require personal finance coursework in the required course denoted in parenthesis. States with listed credit requirement require the denoted number of credits in a standalone required personal finance course. States with a choice of Economics or Personal Finance have personal finance course standards in both courses.

In addition to the College Scorecard, I use the national rollout of personal finance education mandates since 1990 from Stoddard and Urban (2019) which is detailed in Table 1.2. They define the effective year of PFL mandate by the first high school graduating class that is bound by a mandate. PFL standards are most often included in other required courses such as Social Studies, Economics, and Math. However, several of the more recently adopting states have started requiring students to complete a standalone course in personal finance.

Lastly, I use data from the Integrated Postsecondary Education Data System (IPEDS) which includes biannual counts of the incoming cohort of students by previous state of

residence for each university. Between 1986 and 1994, these data were collected every two years from each university and after 1994, universities could voluntarily provide these data to IPEDS in odd years but were required to submit in even numbered years. I use counts of first-time degree seeking students who graduated high school within 12 months of entering college. I replace missing student counts in odd years with linearly interpolated values from neighboring even years.[24]

Additional information on these data sources along with supplemental data sources are detailed in Section 1.8.2.

1.5 Empirical Strategy

1.5.1 Identification

To motivate the empirical strategy introduced in the next section suppose a researcher is able to randomize across a population of N students whether student i will be required to be exposed to PFL topics during high school where $D_i = 1$ denotes those randomly assigned to the mandate and $D_i = 0$ otherwise. After high school, the researcher is able to track an outcome, y_i, for each student i. Due to the random allocation of D_i, the researcher can estimate the causal effect of personal finance education by comparing outcomes across $D_i = 0$ and $D_i = 1$ using the regression specification

$$y_i = \alpha + \gamma^{RCT} D_i + \varepsilon_i$$

where γ^{RCT} is the estimate of the Average Treatment Effect (ATE) of mandated personal finance education on outcome y.[25] Due to the infeasibility of an intervention of this type, a second-best alternative to estimate the causal effect is to exploit a natural experiment in which state policy changes divide the population into mandated students and non-mandated students. Outcomes are then compared across the two populations. Under this difference-in-differences framework, the estimating equation then becomes

$$y_{ist} = \alpha + \gamma^{DD} D_{st} + \delta_{st} + \varepsilon_{ist}, \tag{1.1}$$

where D_{st} now denotes whether state s had a binding mandate for cohort t and δ_{st} is a state-by-year fixed effect. y_{ist} denotes an outcome variable for individual i belonging to graduating cohort t from state s.[26] The outcome variable can be rewritten using the

[24] In Section 1.8.4, I instrument student counts using a combination of fixed effects, observable policy changes, and linear and quadratic trends to replace missing values with estimated values.

[25] Since we have randomization across D_i, we have $E(\varepsilon_i|D_i) = 0$.

[26] In this example, the data are repeated cross section and an individual i is unique to a cohort t and y_{ist}

potential outcomes framework so that

$$y_{ist} = y_{1,ist} \cdot D_{st} + y_{0,ist} \cdot (1 - D_{st})$$

where $y_{1,ist}$ denotes the outcome for an individual if they are bound by a state mandate and $y_{0,ist}$ denotes the outcome if the same individual is not bound by a state mandate. In reality, the researcher only observes either $y_{1,ist}$ or $y_{0,ist}$ for any given individual. However, if the researcher assumes students in states not bound by a state mandate evolve similarly to the unobserved non-mandated students in mandated states, γ^{DD} can be interpreted as the Average Treatment Effect on the Treated (ATT). More formally, suppose there are only two cohorts ($t = 0, 1$) and the state adopting a mandate adopts for the second cohort ($t = 1$). The requisite Parallel Trends Assumption states that

$$E[y_{0,is1} - y_{0,is0} \mid D_{s1} = 0] = E[y_{0,is1} - y_{0,is0} \mid D_{s1} = 1].$$

Under this assumption, observed outcomes for students in the non-adopting states are used as the unobserved counter-factual outcomes for students in the adopting states and the parameter γ^{DD} captures the impact of the state adopted personal finance education on the outcomes for the students who were treated.

However, since micro-level data of this type is not available, consider the case where outcomes are only observed at the university-level for university j. Suppose there exists a function G that maps each student $i \in \mathcal{I}$ to a university $j \in \mathcal{J}$.[27] The outcome $Y_{j\tau}$ is defined as

$$Y_{j\tau} \coloneqq \frac{1}{|J_{j\tau}|} \sum_{i \in J_{j\tau}} y_{ist}$$

where $J_{j\tau}$ is the set students that attend school j, $|J_{j\tau}|$ is the number of students in the set $J_{j\tau}$, and $\tau = t + k_i$ for some k_i which defines the number of periods between graduating high school and appearing in the university-level outcome for student i. Under the enumerated assumptions below, I show in Section 1.8.1 that the parameter γ^{DD} can be consistently estimated using the aggregated estimating equation:

$$Y_{j\tau} = \alpha + \gamma^{DD} \text{pctBound}_{j\tau} + \delta_j + \delta_\tau + e_{j\tau} \tag{1.2}$$

where the term attached to γ^{DD}, $\text{pctBound}_{j\tau}$, corresponds to the fraction of the cohort

is only observed once per individual.

[27] For example, the universities can be indexed such that $\mathcal{J} = \{0, 1, .., J\}$ such that $j = 0$ denotes no university attendance and $j = 1, ..., J$ denotes university attendance.

τ for university j that were bound by state personal finance education mandates. The necessary assumptions for identification are:

1. Parallel Trends Assumption: $E[\Delta y_{0,ist} \mid D_{s(i)t} = 0] = E[\Delta y_{0,ist} \mid D_{s(i)t} = 1] \ \forall t$
2. Cohort Matching Assumption: $k_i = k \ \forall i$
3. Stability of University Mapping: $G(\ i, D_{s(i)t} = 1) = G(\ i, D_{s(i)t} = 0)$

As discussed above, we require the Parallel Trends Assumption in order to satisfy the micro-level difference-in-differences identification strategy. Next, it should be the case that for all i, $k_i = k$. If students in the same high school cohort t enter into different university repayment cohorts τ then it is possible that PFL mandates begin affecting $Y_{j\tau}$ prior to the first mandated cohort as matched by k. This leakage of treated high school cohorts into untreated repayment cohorts will bias the estimate of γ^{DD}. Lastly, it must be the case that assignment of D_{st} does not change the choice of university for students. If students respond to personal finance education by altering the assignment to $J_{j\tau}$, then the estimate from the aggregated specification captures γ^{DD} *plus any potential compositional change* in $J_{j\tau}$ that might affect $Y_{j\tau}$.

Under these assumptions, we have that Equation (1.2) consistently estimates γ^{DD} which is the causal ATE estimated from the micro-level difference-in-differences specification. Section 1.6.2 presents evidence in support of these assumptions and tests the robustness of the results to a loosening of assumptions.

1.5.2 Dose Response Specification

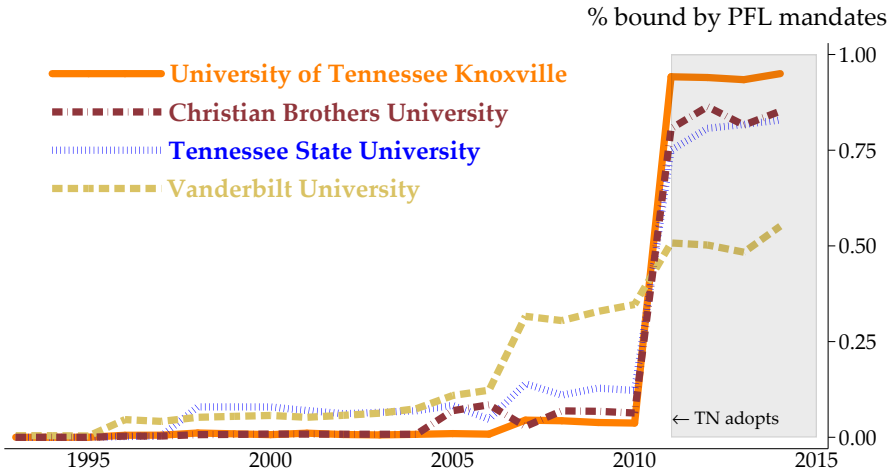
To implement the proposed specification above, I begin by quantifying the share of each incoming cohort bound by PFL mandates in each cohort. Variation at the university-level is driven by two components. The first is through the state adoption of course mandates. When a state changes course standards for high school graduation, all future high school students within the state are affected by this change. When these students graduate high school and proceed to college, universities that receive these students are now populated by these affected students. This is most often universities within the adopting state, however across-state migration of high school students to colleges allows for spillovers from adopting states to non-adopting states. In addition, students from non-adopting states drive down the exposure at colleges in adopting states.

I use the IPEDS previous state of residence data to track within and across-state migration of high school students to colleges. For each university i and incoming cohort t, I construct pctBound_{it} by interacting the state-by-year mandate status of state j for cohort t

(pfMandate$_{jt}$) with the number of students attending university i in cohort t from state j (enroll$_{ijt}$). The total number of mandated students is then divided by the total incoming cohort count from all 50 states and D.C. for cohort t:

$$\text{pctBound}_{it} = \frac{\sum_{j=1}^{51} \text{pfMandate}_{jt} \times \text{enroll}_{ijt}}{\sum_{j=1}^{51} \text{enroll}_{ijt}}. \tag{1.3}$$

Figure 1.3: Examples of Within State Variation in pctBound from Tennessee

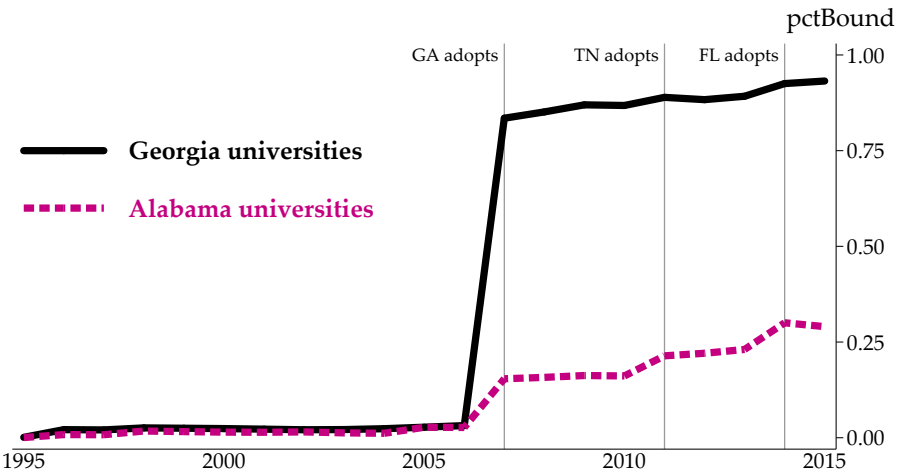


The figure above plots pctBound$_{it}$ as constructed in Equation (1.3) for each university over time. Data on previous state of residence come from IPEDS. State of residence data that are not submitted to IPEDS in odd years are interpolated linearly from neighboring even years. The state of Tennessee adopted a personal finance mandate that was first binding for the class of 2011 as shown by the shaded region in each plot.

Figure 1.3 shows an example of how pctBound$_{it}$ evolves over time for a select group of Tennessee universities. The first graduating class bound by Tennessee's personal finance mandate was the class of 2011. Typically public universities receive a large share of their student body from within the state. However, public universities like University of Tennessee Knoxville and Tennessee State University still differ in the share of students from within-state which drives variation in pctBound after a state adopts a PFL mandate. This heterogeneous impact is even more stark for private universities like Vanderbilt University and Christian Brothers University. Despite the private status, the impact of Tennessee's mandate adoption is larger for Christian Brothers University than for Tennessee State University while Vanderbilt University experiences a smaller shock to pctBound after

2011.

Figure 1.4: Example of Across State Spillovers in pctBound



The figure above plots the state level equivalent of pctBound_{it} where all universities within a state are aggregated. Data on previous state of residence come from IPEDS. Georgia's mandate was first binding for the class of 2007. Alabama did not a adopt mandate during the sample period but is affected by Georgia's adoption through students from Georgia high schools attending Alabama universities and again by Tennessee and Florida's adoption.

In addition to the within-state variation, high school students attending college in other states cause non-adopting states to be affected by mandates in adopting states. Figure 1.4 shows the state level equivalent of pctBound_{it} where the student bodies of all universities within a state are aggregated. When Georgia adopted a mandate binding for the class of 2007, Alabama universities experienced a corresponding increase in pctBound_{it} due to Georgia's adoption. Additionally, Alabama universities experienced subsequent increases in pctBound when Tennessee and Florida adopted in 2011 and 2014, respectively.

I exploit this unique source of exogenous variation in personal finance education exposure to estimate the effect of changes in pctBound_{it} on university-level student loan repayment outcomes. The main specification for the dose response model as motivated by Equation (1.2) is:

$$y_{is,t+k} = \gamma \text{pctBound}_{it} + \beta \mathbf{X}_{ist} + \delta_i + \delta_t + v_{ist}, \tag{1.4}$$

where $y_{is,t+k}$ is an outcome for university i located in state s for high school cohort t and k is the number of periods between cohort t entering college and outcome y being observed. The coefficient of interest is γ which estimates the causal effect of increasing pctBound

from zero to one.

A vector of control variables are also included in $\mathbf{X}_{ist}$ to control for other state level changes that might also affect federal student loan repayment. First, I include controls for the number of credit hours required for high school graduation for math, science, English, and social studies along with the total number of credit hours required. Since the introduction of PFL state standards might be introduced at the same time as changes to other course standards, these controls insure γ is not capturing the effect of changes to other course requirements. Next, I include state level counts of high school staffing for teachers, support staff, and guidance counselors. Changes to state standards might also be accompanied by other state legislation that could increase students' access to college counseling or change student-to-teacher ratios. Without these controls, γ can be biased upward as a result of omitted variable bias. I also include controls for whether cohorts had access to state merit aid scholarships since these may affect where students choose to attend college and how much students pay to attend college. Lastly, I include a vector of unemployment rates between periods t to $t + k$ to control for the local labor market students face during college and into loan repayment. The data sources and construction of these variables are detailed more thoroughly in Section 1.8.2.

Following Assumption 2 above, I match high school graduating cohorts to university repayment cohorts by assuming that students enter repayment after their fourth year of college. Under this assumption, a student graduating high school in year t will enter repayment in fiscal year $t + 5$ and first enter the College Scorecard database in year $t + 6$. As such, all specifications will assume that $k = 6$ for repayment outcomes and $k = 4$ for student loan debt upon entering repayment. I present evidence in support of this assumption along with tests of the sensitivity to the assumptions in Section 1.6.2.

1.5.3 Inference

It is likely that universities within the same state experience common unobserved shocks. Therefore, in the baseline specification, I cluster standard errors at the state level to allow for correlation in the error term, ε_{ist}, between universities in the same state s. However, since treated students are migrating across states to attend college, it is likely that universities in different states also experience common unobserved shocks. If this is the case, errors might be correlated for universities *across states* and, consequently, clustering at the state level may produce standard errors that are too small (Barrios, Diamond, Imbens, and Kolesár, 2012). It is common practice in the treatment effects literature to cluster standard errors at the level of treatment (Bertrand, Duflo, and Mullainathan, 2004;

Cameron and Miller, 2015). In this setting, it is not straight-forward to define the "level of treatment" because each college has a different level of exposure to treatment from various states.

To address this concern, I conduct a randomization inference exercise in the spirit of MacKinnon and Webb (forthcoming). I estimate several "placebo" replications of Equation (1.4) where I randomly draw pfMandate_{jt}. In each replication, I construct placebo pctBound_{it} using the randomly drawn pfMandate_{jt} and the observed enroll_{ijt}. I then estimate Equation (1.4) using the placebo pctBound_{it} to generate placebo $\hat{\gamma}$ estimates. Since the states adopting mandates in the placebo replications are drawn randomly, it must be the case that the γ estimates from this exercise equal zero on average. If $\hat{\gamma}$ estimated using the observed pctBound measure is a sufficiently extreme value in the distribution of placebo estimates, the null hypothesis of no treatment effect can be rejected.

The empirical p-values generated in this algorithm use the distribution of estimated γ coefficients without regard to the standard errors or any assumptions about the correlation structure of the data generating process. Instead, the underlying data generating process of students migrating to universities is captured in the empirical distribution of γ estimates. As a result, the empirical p-values are robust to both within- and across-state correlation of universities driven by enroll_{ijt}. This algorithm is detailed in its entirety in Section 1.8.3.

1.6 Results

1.6.1 Dose Response Specification

The results from the estimation of Equation (1.4) are presented in Table 1.3. Column 1 reports the estimates for the two-year cohort default rate while Columns 2 through 6 report the estimates for the various subsamples of the repayment cohort for the one-year repayment rate. The effect of personal finance mandates on defaults suggests a reduction of 0.2 percentage points (a 5% reduction from the mean) associated with a full dose treatment of the incoming cohort. The magnitude of this estimate is economically meaningful but is not statistically different from zero at conventional levels.[28]

Columns 3 and 4 suggest that increased exposure to personal finance education mandates improves the one-year repayment rate for both first generation and low income students. The point estimates translate to a 2.4 and 2.3 percentage point increase in the repayment rate which corresponds to improvements of 4.5% and 5.1% for first generation and low income students, respectively. Both results are significant at the 5% level regard-

[28]The results are similar for the default rate if the sample years are limited to match the data availability years of the repayment rate analysis.

Table 1.3: Dose Response Estimates: Cohort Default Rate and Repayment Rate

	Default Rate	Repayment Rate				
	(1) Overall	(2) Overall	(3) First Gen	(4) Low Income	(5) Middle Income	(6) High Income
pctBound	-0.002	0.013	0.024	0.023	0.009	0.010
	(0.447)	(0.168)	(0.029)**	(0.034)**	(0.344)	(0.348)
	[0.500]	[0.219]	[0.032]**	[0.037]**	[0.364]	[0.456]
Universities	1,386	1,384	1,340	1,354	1,319	1,317
Cohorts	1993-2006	2001-2008	2001-2008	2001-2008	2001-2008	2001-2008
Outcome Mean	0.042	0.594	0.536	0.458	0.631	0.749
Percentage Effect	-5.0%	2.2%	4.5%	5.1%	1.5%	1.3%

Regressions are weighted using the number of students used to compute each outcome metric. Each column reports a coefficient from a separate regression where the independent variable is pctBound and the outcome is denoted in the column header. The sample includes four-year universities. Default rate analysis includes high school graduating classes 1993 through 2006 and repayment rate analysis includes high school graduating classes 2001 through 2008 due to data availability. First Gen students are defined as students whose parents did not have a college degree. Low Income, Middle Income, and High Income students are defined as household income less than 30,000, between 30,000 and 75,000 and above 75,000, respectively. Controls include cohort weighted credit requirements in math, English, social studies, and science by high school graduation cohort and controls for state level high school staffing, and availability of merit aid scholarships. Also included are university and high school graduation year fixed effects. P-values using standard errors clustered at the state level are presented in parenthesis. Empirical p-values using randomization inference are presented in brackets. *** p<0.01, ** p<0.05, * p<0.1

less of using clustered standard errors (presented in parentheses) or RI-β randomization inference (presented in brackets) to generate p-values. Further, the p-values using both methods are remarkably similar. The point estimates for the overall repayment cohort and for middle and high income students are all positive, but are not statistically different from zero using either p-values. This pattern suggests that personal finance education mandates improve the one-year repayment rate for first generation and low income students.

Table 1.4 repeats the estimation separately for public and private universities in Panel A and Panel B, respectively. The qualitative results are largely unchanged when moving from the full sample in Table 1.3 to the public university sample in Table 1.4. First generation and low income students at public universities have higher repayment rates as a result of PFL mandates. The point estimate and proportional impact for the cohort default rate is again negative but the estimate remains imprecise. In contrast to the results above, high income students at public universities experience a 2.8% improvement in the

repayment rate as a result of PFL mandates which is significant at the 10% level using CRVE and at the 5% using RI-β p-values. This result is contrary to several findings in the literature which find little to no effect of course mandates for students from more affluent backgrounds (Stoddard and Urban, 2019; Goodman, 2019). It is possible that one or more of the mechanisms that cause the improvement in repayment rates operates in a different manner for low income students than for high income students.

The results presented in Panel B suggest there is no significant impact of personal finance education mandates for the sample of private universities. It is possible that the smaller and more frequent shocks to pctBound experienced by many private schools cause a loss of precision in the estimation of the treatment effect. Additionally, it may be the case that some states do not require private high school students to adhere to PFL mandates and these students also attend private universities. The estimated treatment effect on repayment rates is smaller than a 2.1% improvement and many point estimates actually suggest worsening outcomes. As a result, the remaining analysis will focus on the sample of public universities.

Table 1.4: Dose Response Estimates: Cohort Default Rate and Repayment Rate for Public and Private Universities

	Default Rate	Repayment Rate				
	(1)	(2)	(3)	(4)	(5)	(6)
A. Public	Overall	Overall	First Gen	Low Income	Middle Income	High Income
pctBound	-0.003	0.017	0.025	0.022	0.011	0.020
	(0.166)	(0.098)*	(0.025)**	(0.031)**	(0.295)	(0.097)*
	[0.429]	[0.058]*	[0.009]***	[0.035]**	[0.223]	[0.023]**
Universities	450	450	449	449	445	445
Cohorts	1993-2006	2001-2008	2001-2008	2001-2008	2001-2008	2001-2008
Outcome Mean	0.046	0.593	0.551	0.478	0.630	0.730
Percentage Effect	-5.9%	2.8%	4.5%	4.7%	1.8%	2.8%
B. Private						
pctBound	0.001	-0.009	0.007	0.009	-0.007	-0.014
	(0.877)	(0.682)	(0.676)	(0.716)	(0.660)	(0.250)
	[0.820]	[0.370]	[0.981]	[0.704]	[0.242]	[0.070]*
Universities	936	934	891	905	874	872
Cohorts	1993-2006	2001-2008	2001-2008	2001-2008	2001-2008	2001-2008
Outcome Mean	0.036	0.595	0.512	0.423	0.632	0.777
Percentage Effect	3.2%	-1.5%	1.3%	2.1%	-1.1%	-1.8%

Regressions are weighted using the number of students used to compute each outcome metric. Each column reports a coefficient from a separate regression where the independent variable is pctBound and the outcome is denoted in the column header. The sample includes public and private four-year universities. Default rate analysis includes high school graduating classes 1993 through 2006 and repayment rate analysis includes high school graduating classes 2001 through 2008 due to data availability. First Gen students are defined as students whose parents did not have a college degree. Low Income, Middle Income, and High Income students are defined as household income less than 30,000, between 30,000 and 75,000 and above 75,000, respectively. Controls include cohort weighted credit requirements in math, English, social studies, and science by high school graduation cohort and controls for state level high school staffing, and availability of merit aid scholarships. Also included are university and high school graduation year fixed effects. P-values using standard errors clustered at the state level are presented in parenthesis. Empirical p-values using randomization inference are presented in brackets. *** p<0.01, ** p<0.05, * p<0.1

1.6.2 Tests of Identification Assumptions

As discussed in Section 1.5.1, the estimates presented above can only be interpreted as the micro-level ATE if Assumptions 1 through 3 are satisfied. This section provides evidence for each assumption.

Cohort Matching Assumption

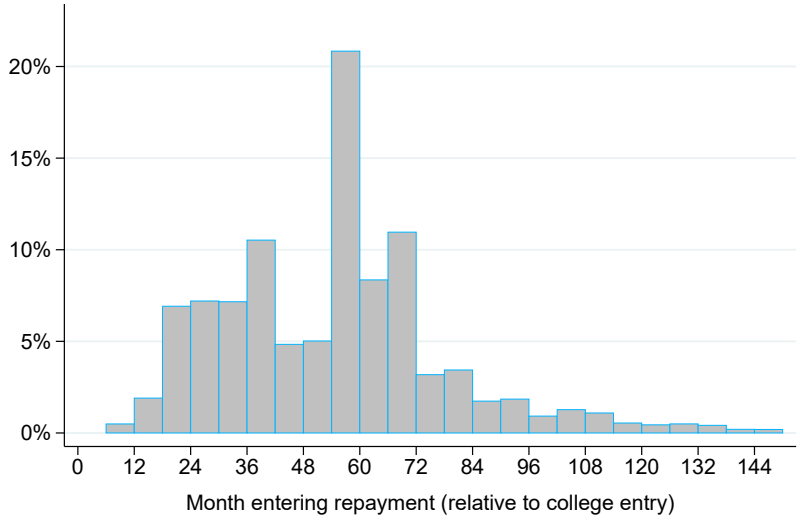
In Figure 1.5, I show the distribution of the month students in the 2003 incoming college cohort entered repayment to support the choice of $k = 6$ as the appropriate lag between entering college and observing repayment outcomes. The model month students entered repayment is 51 months after entering college. This timing is consistent with a student entering college in August, graduating in May of their fourth year, and entering repayment in November. From this data, 43% of the repayment cohort enter repayment between 48 months and 72 months after entering college. However, the data show a significant percentage of students entering repayment prior to 42 months since entering college. If these are students bound by personal finance education mandates, it is possible they contribute to a repayment cohort that is inconsistent with the assumption of $k = 6$. The specification in the next section will test whether students separating from college prior to four years impact the repayment rate for a university.

Parallel Trends Assumption

Identification of the causal effect of PFL education on student loan repayment outcomes relies on the Parallel Trends Assumption. This assumption is not directly testable since counter-factual outcomes for treated units are unobserved. Instead, I present evidence that universities more exposed to PFL mandates were not trending differentially prior to the adoption of mandates by estimating a flexible event study specification. The event study specification includes a vector of binary variables in which each variable represents a time period relative to the start of treatment for units experiencing an event. Each parameter then estimates the difference in outcomes between units experiencing an event and units not experiencing an event in the time period relative to treatment. In this context, the treatment variable is not discrete and thus care must be taken to define the first period of treatment. I define a university event as a year-over-year change in pctBound_{it} of 50 percentage points or larger:

$$\text{event}_{ist} = 1 \cdot \{\text{pctBound}_{it} - \text{pctBound}_{i,t-1} \geq 0.5\}. \tag{1.5}$$

Figure 1.5: Month Entering Repayment for 2003 High School Cohort



The figure above plots a histogram of the month a borrower enters repayment relative to the month they enter college for four-year college students who did not attend graduate school. The sample includes federal student loan borrowers from the cohort entering college in the 2003-2004 academic year. Students who attended graduate school are removed since they would mechanically enter into repayment at a later month. Total student counts are collapsed into six month bins and nationally representative weights are used to create cohort shares.
Source: U.S Department of Education, National Center for Education Statistics, 2004/2009 Beginning Postsecondary Students Longitudinal Study Restricted-Use Data File.

I choose the 50 percentage point threshold to ensure a university can only experience one event. Table 1.13 details the number of universities experiencing an event by this definition in each academic year along with the states adopting in each year. Between 1996 and 2014, 524 universities experience an academic year in which the adoption of at least one mandate changes pctBound_{it} by at least 50 percentage points. Although there are a few early adopting states, most of the events occur for the high school graduating cohorts of 2005 and later. As a result, there is more outcome data available for the periods prior to an event than for the periods after an event. The estimating equation for the event study is identical to Equation (1.4) aside from the event study parameters:[29]

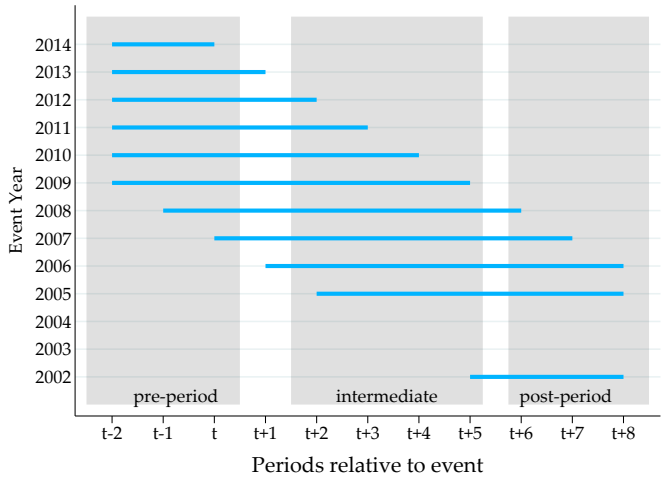
$$y_{is\tau} = \sum_{j=-2}^{0} \gamma_j \text{event}_{is,t+j} + \sum_{j=2}^{10} \gamma_j \text{event}_{is,t+j} + \beta \mathbf{X}_{ist} + \delta_i + \delta_\tau + \varepsilon_{is\tau}. \tag{1.6}$$

In this specification, the event occurs in period t and a separate parameter is estimated for each period relative to an event with period $t+1$ being omitted.[30] This results in ten

[29]The vector of control variables $\mathbf{X}_{is,t+6}$ uses the assumed matching high school cohort corresponding to repayment outcome $y_{is\tau}$.

[30]Since it takes at least one year for the repayment rate outcome to be observed, period $t+1$ is the last

Figure 1.6: Contribution to Event Study Parameters by University Event Year



The above graph shows how each university contributes to each event study parameter by the year of university event. Universities experiencing an event in 2009 and later do not contribute to the post-event parameters while universities experiencing an event before 2007 do not contribute to the pre-event parameters. States adopting between 2006 and 2008 contribute to both the pre-period and post-period parameters.

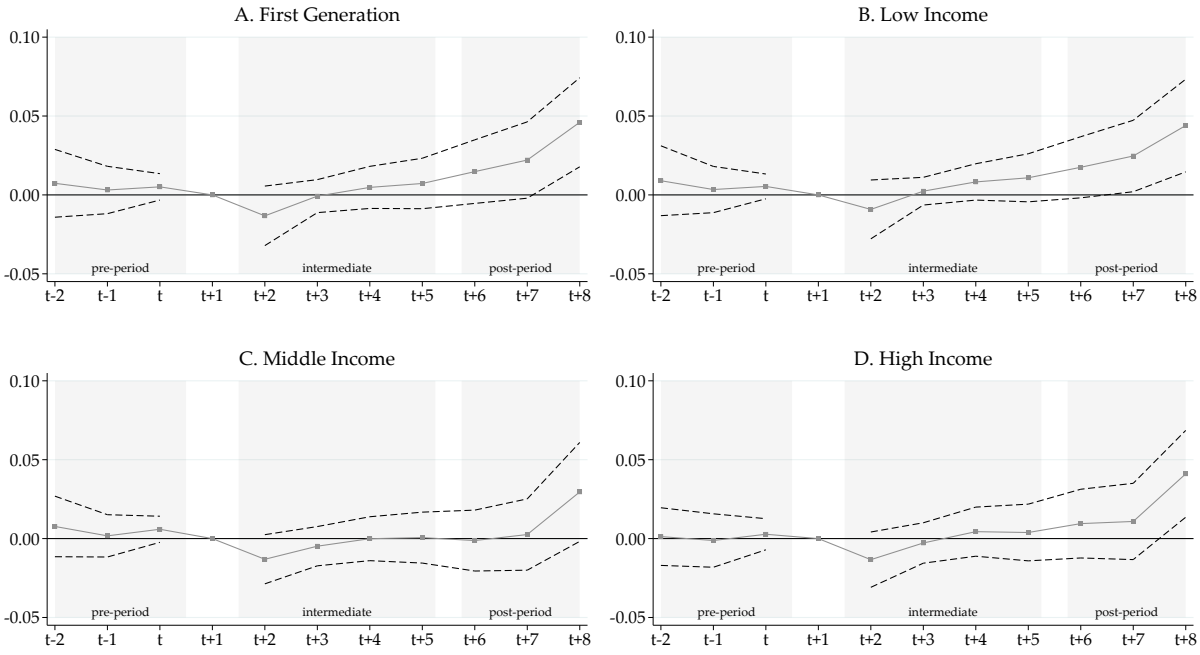
estimated parameters across the event space. Since the College Scorecard only contains eight years of data for the repayment rate, the window for each university's outcome will not span the entirety of the range of event study parameters. Hence, the event study coefficients represent a combination of the dynamic effect of each university's change in outcomes over time plus a heterogeneous effect of universities entering and exiting the identification of the parameter space. This is more clearly shown in Figure 1.6 which plots the identifying variation of each event across the event study parameters. Since the data is both right and left censored, the universities that identify the pre-period (periods $t-2$ through period t) are largely universities located in the states adopting after the 2008 high school graduating class. On the other hand, the universities identifying the post-period (periods $t+6$ through $t+8$) are the states adopting in 2008 and prior. The parameters corresponding to periods $t+2$ through $t+5$ represent an intermediate range in which it is possible that treated students enter the Scorecard data if they leave college prior to the assumed four year spell. This intermediate range of parameters tests for changes to repayment outcomes as a result of non-completing treated students.

Figure 1.7 reports the coefficients and 95% confidence intervals for the estimation of Equation (1.6) for the repayment rate subgroups at public universities. In all four panels, there is no differential trend between universities experiencing an event and those not

period before treated students can begin contributing to a university's repayment rate for each university. A treated student entering the repayment rate data in year $t+1$ would be a student who separated from college prior to the end of the first year.

experiencing an event. Further, there does not appear to be a significant effect on university repayment rates during the intermediate periods for any of the subsamples. The estimated coefficient for period $t+2$ is negative for all groups and represents the smallest parameter estimate but it rather small in magnitude. Although none of the intermediate estimates is distinguishable from zero, the upward trend suggests non-completers exposed to personal finance mandates may also be better at repaying student loans. The estimates for the post-period parameters for first generation (low income) students range from 1.5 percentage points (1.1 percentage points) in period $t+6$ to 4.6 percentage points (4.4 percentage points) in period $t+8$ which is consistent with the point estimates presented in the main specification. The results in Panels C and D for middle and high income students suggests no significant effect until period $t+8$ and is largely consistent with the muted estimates presented in Table 1.4.

Figure 1.7: Event Study Coefficients for Sample of Public Universities



Each panel in the above figure presented the vector of event study parameters with period $t+1$ omitted as the reference period. Since the repayment rate data takes at fewest one year to enter the College Scorecard, it is not possible for a member of high school cohort t to contribute to the repayment outcome in $t+1$. However, early separators can contribute to the parameters $t+2$ through $t+5$. Period $t+6$ represents students who spend four years in college and periods greater than $t+6$ represent students from cohort t who spent longer than four years in school or students in cohorts greater than t which were also bound by PFL mandates.

In total, these results present evidence in support of the Parallel Trends Assumption

and in support of the Cohort Matching Assumption. In each event study specification, the coefficients corresponding to periods prior to event are both small and indistinguishable from zero. The same is true for each of the parameters corresponding to periods in which treated non-completers might contribute to the university repayment rate. Lastly, the parameters corresponding to the post-periods suggest that first generation and low income students have higher repayment rates. The point estimates suggest that when the third cohort after the event has spent four years in college, repayment rates for first generation and low income students are 4.6 and 4.4 percentage points higher, respectively.

Stability of University Mapping Assumption

In this section, I present evidence that PFL mandates do not cause students to alter their college enrollment decisions. In order for university-level outcomes to appropriately proxy aggregated micro-level outcomes, it must be the case that exposure to PFL education does not alter the university a student chooses to attend. If exposure to required PFL coursework shifts the mapping of students to universities, changes in federal student loan repayment outcomes at the university-level might be due to compositional shifts in the student body as a result of PFL mandates. In this case, it is possible to detect improvements in university-level repayment outcomes without any micro-level improvements.

To test this assumption, I use the IPEDS previous state of residence data to track the flow of high school students from each state into the colleges they ultimately enroll.[31] Recall the IPEDS data includes the variable enroll_{ijt} which is the number of students in the incoming cohort for university i who previously resided in state j for incoming cohort t. Instead of aggregating student counts at the university-level, I can instead aggregate student counts at the previous state of residence level according to universities characteristics. This procedure generates a measure of the percentage of high school students from each state attending universities of a given type. Equation (1.7) illustrates an example of this variable construction using the university characteristic Public4yr_i, which equals one if a university is a four-year public college and Seniors_{jt} is the total number of enrolled high school seniors for the graduating cohort t.[32]

$$\text{pctPublic4yr}_{jt} = \frac{\sum_{i \in I} \text{Public4yr}_i \times \text{enroll}_{ijt}}{\text{Seniors}_{jt}} \tag{1.7}$$

[31] Stoddard and Urban (2019) perform a similar test using the IPEDS enrollment data and counts of the number of 18 year olds in a state in a given year.

[32] The variable Seniors_{jt} comes from enrollment counts from the Department of Education's Common Core of Data.

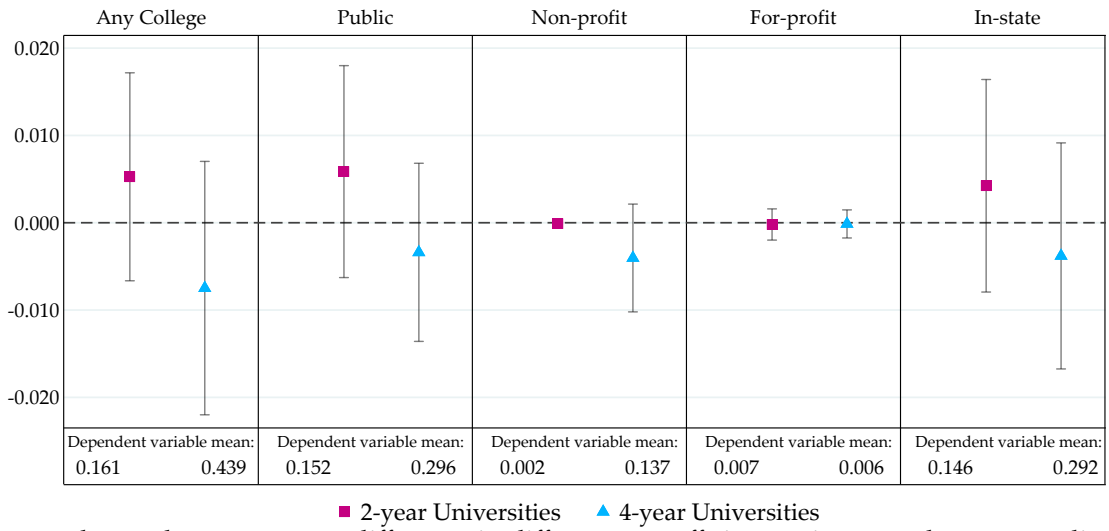
The constructed variable, pctPublic4yr$_{jt}$ measures the percentage of high school seniors from state j in cohort t who enrolled in a public four-year university. I create analogous variables for any two-year and four-year college and for two-year and four-year public, private non-profit, private for-profit, and in-state universities.

I use these constructed variables as outcome measures for a state-by-year difference-in-differences design to test whether the state adoption of a personal finance education mandate alters the share of students attending various types of colleges. The estimating equation is similar to Equation (1.4):

$$y_{jt} = \gamma \text{pfMandate}_{jt} + \beta \mathbf{X}_{jt} + \delta_j + \delta_t + \varepsilon_{jt}, \tag{1.8}$$

where y_{jt} is a state-level variable, such as pctPublic4yr$_{jt}$, and pfMandate$_{jt}$ is a binary variable denoting whether state j had a binding personal finance mandate in effect for cohort t. The vector X_{jt} is the same set of state-by-graduation year level controls as in Equation (1.4) and δ_j and δ_t are state and year fixed effects, respectively.[33]

Figure 1.8: Difference-in-Differences Estimates for Changes to College Enrollment



The figure above plots a separate difference-in-differences coefficient estimate and corresponding 95% confidence interval where the independent variable is pfMandate and the outcome variable is denoted for each column. Each outcome is reported for two-year universities with a square and four-year universities with a triangle. Control variables include state level counts of high school staffing, other high school graduation credit requirements for math, English, social studies, and science, and the availability of state merit aid scholarships.

Figure 1.8 presents point estimates and 95% confidence intervals for this specifica-

[33]These controls include state level counts of school staffing, other course requirements for graduation,

tion.[34] All reported point estimates a smaller than a one percentage point change in either direction and no estimate is statistically different from zero at any conventional significance threshold. The results are consistent with the findings in the literature which find no changes in college attendance or the choice of college as a result of a binding personal finance education mandate (Stoddard and Urban, 2019).

Robustness to Changes in College Enrollment

In this section, I present evidence that the results presented above are robust even in the case of changes to student enrollment decisions. I estimate a variation of Equation (1.4) using an alternate construction of pctBound. This specification is motivated by a "shift-share" framework in which exposure levels are held constant at their initial levels and the variation in the identifying variable is driven by an interaction of the constant shares and an aggregate trend (Bartik, 1987).[35] Hence, the identifying variation in this model does not rely on transitory changes in high school students' college choice but rather each university's exposure to each state's potential adoption of PFL mandates in the period before PFL mandate adoption. Applying this framework to the construction of pctBound, I construct $\overline{\text{enroll}}_{ij,S}$ which is the mean enrollment of the students from state j at university i for a set of academic years S. The construction of $\widehat{\text{pctBound}}_{it}$ takes the form:

$$\widehat{\text{pctBound}}_{it} = \sum_{j=1}^{51} \left[\frac{\overline{\text{enroll}}_{ij,S}}{\overline{\text{enroll}}_{i,S}} \right] \times \text{pfMandate}_{jt} \quad . \tag{1.9}$$

For this analysis, the set S contains the IPEDS state of residence counts from 1986 through 1994 as this period largely contains state composition before the rollout of personal finance mandates. Summing over all states and D.C. yields the mean total enrollment $\overline{\text{enroll}}_{i,S}$ for university i during the set of years S. Hence, the fixed share of students from university i from state j can be derived as the ratio of $\overline{\text{enroll}}_{ij,S}$ to $\overline{\text{enroll}}_{i,S}$. When there is no change in PFL mandate adoption from year t to year $t+1$, there is no change in $\widehat{\text{pctBound}}_{it}$ to $\widehat{\text{pctBound}}_{i,t+1}$. However, if state J adopts a mandate between graduating cohort t and graduating cohort $t+1$, the difference in $\widehat{\text{pctBound}}_{i,t+1}$ and $\widehat{\text{pctBound}}_{it}$ is exactly equal to the ratio $\overline{\text{enroll}}_{iJ,S}/\overline{\text{enroll}}_{i,S}$, or state J's historical composition for university i.

Figure 1.9 plots the estimates from this alternative specification relative to the estimates

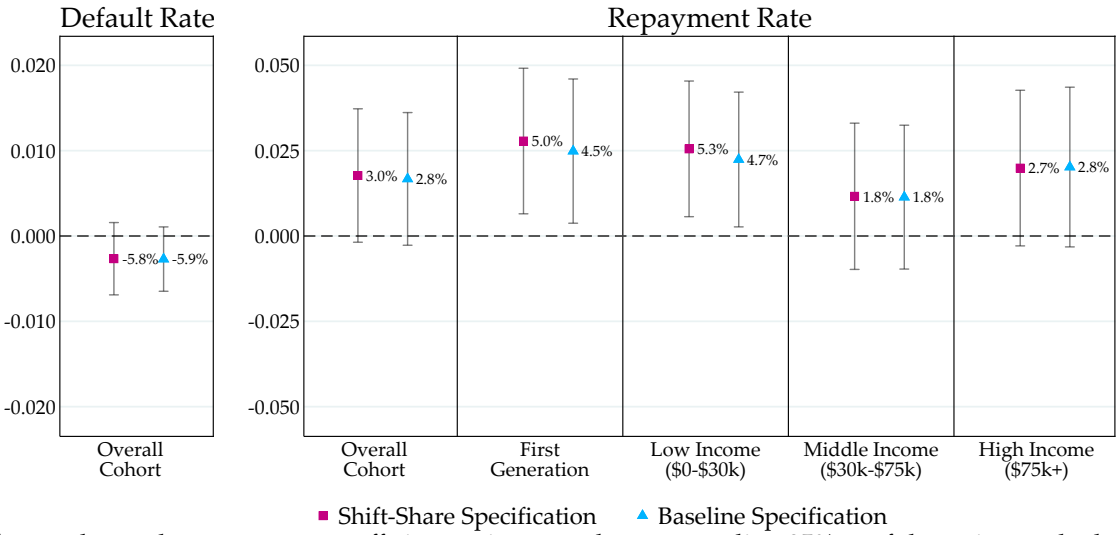
and the availability of state merit aid scholarships.

[34]The full table of results are available in Tables 1.14 and 1.15.

[35]Recall in Section 1.8.1 that identification of γ^{DD} requires that D_{st} does not affect college choice and that the reweighted state fixed effects by the state share but be included in the error term. However, this specification holds state shares fixed which allows for γ^{DD} to be consistently estimated without this assumption.

from the baseline specification for the sample of public universities. For all outcome variables, the point estimates are largely unchanged between the baseline specification and the shift-share specification. In fact, the point estimates using the shift-share specification represent larger impacts for both first generation and low income students. These results confirm that the improvements in PFL mandates stem from a university's exposure to PFL mandates rather than from transitory changes in student enrollment induced by PFL mandates.

Figure 1.9: Dose Response Shift-Share Estimates Relative to Baseline Specification for Public Universities



The figure above plots a separate coefficient estimate and corresponding 95% confidence interval where the independent variable is pfMandate and the outcome variable is denoted for each column. The estimates from the Shift-Share specification are reported with a square and the estimates from the baseline specification are replicated with a triangle. Proportional impacts are also printed to the right of each marker. The vector of control variables remains unchanged from the main specification. 95% confidence intervals are constructed using standard errors clustered at the state level.

1.6.3 Application: Universal PFL Mandates

To illustrate the magnitude of these estimates, I construct a counter-factual using the results from the baseline specification. Consider a hypothetical case in which all high school students graduating between 2001 and 2008 were bound by state PFL mandates. Table 1.5 shows the total pctBound across all public universities for the incoming cohort matched to each cohort's one-year repayment rate. I also present the total number of

[36]Since the repayment rate is calculated using two cohorts, I divide the repayment cohort in half.

students in each repayment cohort that successfully repaid at least one dollar of their loan balance one year after entering repayment.[36] The next panel shows how the one-year repayment rate changes when I apply the estimated ATT for the overall cohort from Table 1.4 scaled by the share of the student body not bound by mandates. I also report the total number of students successfully repaying federal students loans under this assumption along with the estimated increase in the number of students meeting this metric. On average, around 9,000 additional students would have paid down at least one dollar of original principal after one year for a total of over 72,000 additional borrowers making progress on their loans.

One important consideration is whether one-year repayment outcomes are predictive of long-term repayment success. Table 1.6 shows a summary of various long term repayment outcomes for a nationally representative sample of college students entering college in 2003. The sample is split by whether a student had paid down at least one dollar in principal one year after entering repayment. Students in this cohort who were able to pay down at least a dollar of their student loan debt one year after entering repayment were significantly better at repaying their loans 12 years after entering repayment. These students paid down nearly half of their loans while those not hitting the benchmark still owed 81% of their original balance. They were also half as likely to have defaulted on a student loan and were 19 percentage points more likely to have ever repair their student loans. Although not causal estimates, these comparisons suggest that improvements in the one-year repayment rate caused by PFL mandates could also lead to large future student loan repayment success for mandated students.

Table 1.5: Hypothetical Effect of Universal Personal Finance Mandate

	Observed			Hypothetical		
Cohort	pctBound	Repayment Rate	Students Repaying	Repayment Rate	Students Repaying	Δ Students Repaying
2001	12.0%	71.7%	418,262	73.1%	426,868	+ 8,606
2002	12.0%	65.1%	333,385	66.6%	341,076	+ 7,691
2003	12.2%	60.9%	325,790	62.4%	333,815	+ 8,025
2004	12.6%	58.5%	342,921	60.0%	351,702	+ 8,781
2005	16.5%	56.5%	372,848	57.9%	382,312	+ 9,464
2006	17.0%	54.6%	413,066	56.0%	423,862	+ 10,796
2007	32.2%	53.4%	442,683	54.5%	452,285	+ 9,602
2008	33.7%	53.9%	460,313	55.0%	469,992	+ 9,679
						+ 72,644

The table above details a hypthetical exercise which assumes the estimated effect of PFL mandates is applied to all unmandated students in each entering cohort. The estimated treatment effect from Table 1.4 for the overall cohort is 0.017. The hypothetical repayment rate is computed by adding $(1 - \text{pctBound}_{ist}) * 0.017$ to the observed repayment rate.

Table 1.6: Long-term Repayment Outcomes Conditional on One Year Repayment (Beginning Postsecondary Students 2004)

	Paid down principal after one year	
Outcome 12 years after entering college	Yes	No
Percent owed on balance	0.51	0.82
Ever defaulted on loan	0.12	0.24
Ever paid off loan	0.58	0.37
Remaining balance	$22,086	$36,814
Total Weighted Population	537,990	425,930

Source: U.S Department of Education, National Center for Education Statistics, 2004/2009 Beginning Postsecondary Students Longitudinal Study Restricted-Use Data File with the 2015 FSA Supplement. Estimates come from author's calculations. One year repayment rate metric is constructed by calculating the outstanding student loan balance one year after entering repayment and comparing to the outstanding balance upon entering repayment. Only borrowers who had entered repayment by the end of 2009 are considered to maintain consistent end dates.

1.6.4 Evidence of Mechanisms

Student Loan Debt

To test for changes in federal student loan borrowing, I estimate Equation (1.4) on moments from the student loan debt distribution and the median debt for subsamples of the student body from the College Scorecard. Table 1.7 reports the estimated effect of personal finance mandates on student loan debt upon entering repayment for public universities. The sample is restricted to the high school graduation years 2001 through 2008 to match the results presented above. The effect of increased exposure to personal finance education on student loan debt is proportionally small for the 10th, 75th, and 90th percentiles of the debt distribution for public universities. However, the estimates at the 25th percentile and the median represent imprecise reductions in student loan debt around 3%.

Table 1.7: Dose Response Estimates: Moments from Student Loan Debt Distribution for Public University Sample

	(1) 10th	(2) 25th	(3) 50th	(4) 75th	(5) 90th
pctBound	-2.8 (50.1)	-122.0 (127.3)	-305.2 (234.6)	-312.7 (269.9)	-144.3 (377.1)
Universities Cohorts	449 2001-2008	450 2001-2008	450 2001-2008	450 2001-2008	449 2001-2008
Outcome Mean Percentage Effect	2480.3 -0.1%	4181.4 -2.9%	9516.4 -3.2%	18172.4 -1.7%	25959.9 -0.6%

Regressions are weighted using the number of students used to compute each outcome metric. Each column reports a coefficient from a separate regression where the independent variable is pctBound and the outcome is denoted in the column header. The sample includes public four-year universities. 10th, 25th, 50th, 75th and 90th each represent the correspondent moment in a unviersity's student loan debt levels for students entering repayment. Controls include cohort weighted credit requirements in math, English, social studies, and science by high school graduation cohort and controls for state level high school staffing, and availability of merit aid scholarships. Also included are university and high school graduation year fixed effects. P-values using standard errors clustered at the state level are presented in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

This effect can be further explored by tracing the effect on the same subsamples of the student body discussed above for the one-year repayment rate. The improvements in repayment rates were largest for low income and first generation students with smaller

Table 1.8: Dose Response Estimates: Median Student Loan Debt by Student Subsample for Public University Sample

	(1) First Gen	(2) Low Income	(3) Middle Income	(4) High Income
pctBound	-272.3	68.2	-281.9	-783.4**
	(247.9)	(210.1)	(232.6)	(317.6)
Universities	449	450	446	446
Cohorts	2001-2008	2001-2008	2001-2008	2001-2008
Outcome Mean	9288.9	9289.7	9911.5	9602.7
Percentage Effect	-2.9%	0.7%	-2.8%	-8.2%

Regressions are weighted using the number of students used to compute each outcome metric. Each column reports a coefficient from a separate regression where the independent variable is pctBound and the outcome is denoted in the column header. The sample includes public four-year universities. First Gen students are defined as students whose parents did not have a college degree. Low Income, Middle Income, and High Income students are defined as household income less than 30,000, between 30,000 and 75,000 and above 75,000, respectively. Controls include cohort weighted credit requirements in math, English, social studies, and science by high school graduation cohort and controls for state level high school staffing, and availability of merit aid scholarships. Also included are university and high school graduation year fixed effects. P-values using standard errors clustered at the state level are presented in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

effects for high income students. Table 1.8 reports the effect of personal finance mandates on the median loan debt for subgroups of the public university cohorts. The point estimates are negative and imprecise for first generation and middle income students. On the other hand, the estimate for high income students is sizable and statistically significant, representing a decline in borrowing of around 8%. This heterogeneous response from personal finance education may help to explain the improvements in repayment for high income students. While there is little evidence of changes to borrowing patterns for first generation and low income students who saw the largest effects on repayment rates, it may be the case that high income students have better repayment rates as a result of lower student loan balances upon entering repayment. This could be due to a decision to borrow less or due to increases in grant or scholarship receipt as found in Stoddard and Urban (2019).

Information Intervention

In this section, I test whether students who were bound by personal finance education mandates in high school are better able to answer questions about financial literacy and federal student loans. I employ a difference-in-differences design with micro-level data

from three nationally representative surveys. The estimating equation for each data source is similar and takes the form

$$y_{isjt} = \gamma \text{pfMandate}_{st} + \beta X_{isjt} + \delta_s + \delta_t + \delta_j + \varepsilon_{isjt}, \tag{1.10}$$

where y_{isjt} is a binary variable for whether respondent i from state s observed in survey wave j graduating from high school in year t correctly answered a particular question. I drop any respondent with a GED or no high school diploma since these students were not bound by state graduation mandates. X_{isjt} includes a vector of binary control variables which includes race, gender, and education and the vector of state-by-graduation year controls for merit aid, high school staffing, and credit requirements as in Equation (1.4). γ is the parameter of interest which estimates the impact of a binding personal finance education mandate on the (linear) probability of correctly answering the question. Lastly, I include state, survey wave, and high school graduation year fixed effects and I use the included state-level survey weights so the analysis is representative of each state's population.[37] Standard errors are clustered at the high school state level.

The first survey is the National Postsecondary Student Aid Study (NPSAS) which includes a nationally representative sample of college students every two years. The 2016 wave of the NPSAS began asking college students three financial literacy questions and three questions pertaining to knowledge of federal student loan repayment.[38] These new data provide insights into the financial literacy of current college students that was previously not available. However, since these questions are only included in one survey wave, comparisons of mandated students to not mandated students from the same state largely rely on students surveyed at different ages. As a result, the estimate for γ potentially includes the effect of personal finance education mandates plus a bias term. Additionally, the 2016 wave of the NPSAS largely contains students that graduated high school after the period between 2001 and 2008 studied above. Regardless, the novelty of the questions asked in this survey necessitate its use.

In addition to the NPSAS, I also use data from the National Financial Capability Study (NFCS) and the Survey of Household Economics and Decisionmaking (SHED). The NFCS data contain waves from 2012, 2015, and 2018 while the SHED data contain waves from 2017 and 2018. Both surveys are nationally representative and each asks five financial literacy questions that largely overlap in content.[39] Since each survey contains multiple waves, it is possible to compare respondents surveyed from the same state and at the

[37]The weights included in the NPSAS:16 are nationally representative instead of state representative.
[38]The questions are detailed in Table 1.20.
[39]Survey question text can be found in Table 1.20.

same age but with different values for pfMandate. However, neither survey includes data on state or year of high school graduation. Rather, I follow the convention in Urban, Schmeiser, Collins, and Brown (2018) and Harvey (2019) and assign mandate status by state of residence and year of 18th birthday. Additionally, for these two surveys, I restrict the sample to only those students whose (inferred) high school graduation year is between 2001 and 2008 (inclusive) in order to match the data years for the improvements in the one-year repayment rate discussed above.

Figure 1.10 plots the coefficient estimates and 95% confidence intervals from the estimation of Equation (1.10) for the three surveys.[40] Panel A plots the coefficient estimates for the three financial literacy and three loan literacy questions from the NPSAS:16. The point estimates for each of the three financial literacy questions is less than a half a percentage point and the null hypothesis cannot be rejected for any estimate. Further, there is no significant change in the total number of correct answers as a result of a binding personal finance mandate.[41] On the other hand, each of the estimates for the three loan literacy questions is positive ranging from 1.3 to 3.3 percentage point increases in the probability of a correct answer. The largest effect is for the question asking borrowers whether the federal government can garnish wages for non-payment of federal student loans. Additionally, respondents answered 3% more questions correctly if they were bound by a personal finance mandate.

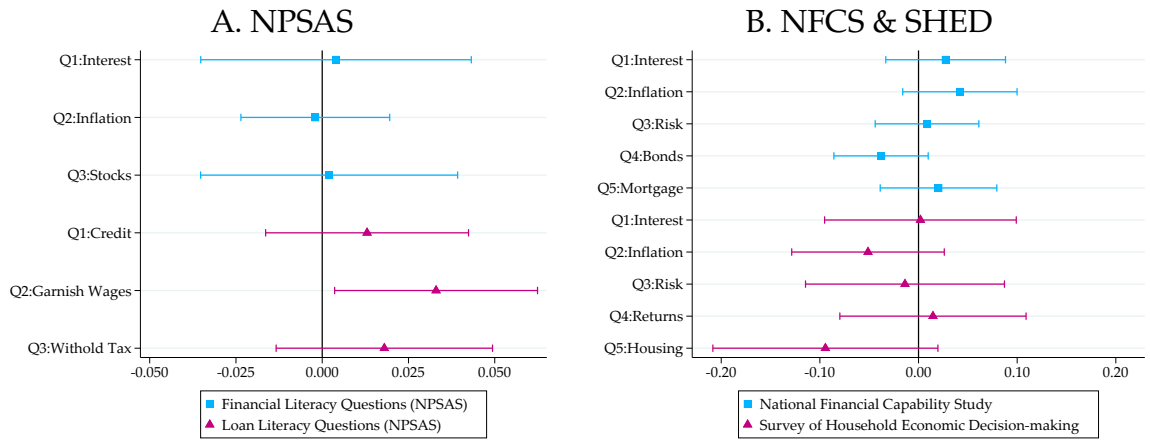
Panel B plots the point estimates for the five financial literacy questions in both the NFCS and the SHED surveys. Across all ten questions, no null hypothesis can be rejected at any conventional level. In total, the evidence suggests no difference in the probability of correctly answering financial literacy questions between mandated and not mandated respondents *at the time of survey*. However, these results should not be taken as evidence that personal finance education mandates do not improve financial literacy. The evidence does not preclude the case where PFL mandates improve financial literacy during and immediately after high school and either financial literacy depreciates quickly or non-mandated peers catch up to mandated peers after high school. If this is the case, PFL mandates may still improve downstream outcomes due to decisions made during high school while financial literacy was higher than non-mandated peers.

On the other hand, the evidence does suggest that mandated students are more knowledgeable about regulations governing federal student loans. If this is the case, student loan borrowers may be better able to repay loans due to this increased familiarity with

[40] Tables 1.17 to 1.19 report the full results for the NPSAS, SHED, and NFCS, respectively.

[41] The full table of coefficients and standard errors along with the effect on the number of correct answers can be found in Table 1.17.

Figure 1.10: Difference-in-Difference Estimates for Financial Literacy and Loan Literacy from NPSAS:16



The above figures plot the point estimates and 95% confidence intervals for the difference-in-difference coefficients using the NPSAS, NFCS, and SHED surveys. Panel A plots the effect of PFL mandates on the linear probability for three financial literacy and three loan literacy questions. Panel B plots the effect of PFL mandates on the linear probability of five financial literacy questions for each survey. Question text is available in Table 1.20 and a full table of coefficient estimates and standard errors is available in Tables 1.18 to 1.19.

Sources: U.S Department of Education, National Center for Education Statistics, Restricted-use National Postsecondary Student Aid Study 2016. Also National Financial Capability Study and Survey of Household Economics and Decisionmaking

the rules and regulations for their loans. These students might be aware of income-driven repayment plans or deferment or forbearance options. In addition, Stoddard and Urban (2019) find that students bound by personal finance education mandates are more likely to borrow from federal sources. It could also be the case that the increase in knowledge about the federal student loan system is due to an increase in the probability of federal borrowing. Anderson, Conzelmann, and Lacy (2018) find that federal borrowers have higher student loan literacy which might be a result of more experience with the federal loan system.

Completion

Lastly, I test whether personal finance education mandates have any impact on the probability a student earns a degree. If personal finance education leads to better matching of students to colleges or degree programs, students may be more successful in college. As a result of graduation, students will likely have better labor market outcomes which would lead to better repayment rates and a lower chance of default. I use the American Community Survey (ACS) one-year samples from 2005-2017 to test whether students bound by personal finance mandates were more likely to hold a college degree or have

ever attended college. The estimating equation for these tests is identical to Equation (1.10). However, I remove students younger than 22 as these students are unlikely to have earned a bachelor's degree yet. Since the ACS also does not ask respondents for the year of high school graduation, I assume respondents graduate from high school in the year of their 18th birthday. I report results where either the state of birth or the state of residence is used in place of state of high school since state of high school graduation is also unobserved.

Table 1.9 reports the results of this estimation on the (linear) probability of earning a bachelor's degree, the probability of earning an associate's degree, and the probability a respondent ever attended college. Odd numbered columns identify mandate status by birth state while even numbered columns use state of residence. The estimate in Column 1 suggests a potential negative relationship between personal finance mandates and bachelor's degree receipt although the point estimate is quite small. However, when using the state of residence instead of birth state in Column 2, the estimate is not statistically significant. The estimates in Columns 3 through 6 suggest there is also no effect of personal finance education mandates on the probability of earning an associates degree or having ever attended college. In total, I find little compelling evidence to suggest that the improvement in student loan repayment is due to an increase in degree completion.

1.7 Conclusion

The findings in this paper extend the literature on personal finance education mandates and federal financial aid in several key dimensions. I find that students who were bound by PFL mandates in high school were better at repaying student loan balances. The impact is largest and most precisely estimated for low income and first generation students at public universities which is consistent with other findings in the literature (Stoddard and Urban, 2019; Goodman, 2019). The results suggest that low income and first generation students are 5% more likely to have paid down some of their original balance one year after entering repayment. Despite some suggestive evidence of improvements, I cannot conclude that mandates have any meaningful impact on the cohort default rate. However, this result is likely not surprising since student loan default is a more rare and adverse outcome while repayment progress is a more sensitive measure.

I conduct a counter-factual exercise to estimate how many additional students would have been able to successfully pay down some of their student loan balance if PFL mandate were universal. If all students were bound by PFL mandates for the high school graduating cohorts between 2001 and 2008, an additional 72,000 students would have paid down at least a dollar of their balance on year after entering repayment. I show evidence that

Table 1.9: Difference-in-Differences Estimates for Degree Completion from ACS

	(1) Bachelor's Earned	(2) Bachelor's Earned	(3) Assoc Earned	(4) Assoc Earned	(5) Ever College	(6) Ever College
PF Mandate	-0.005*	-0.004	0.001	-0.002	-0.007	-0.006
	(0.003)	(0.003)	(0.001)	(0.002)	(0.004)	(0.005)
Observations	2,236,990	2,236,990	2,236,990	2,236,990	2,236,990	2,236,990
Cohorts	2001-2008	2001-2008	2001-2008	2001-2008	2001-2008	2001-2008
High School State	Birthplace	Residence	Birthplace	Residence	Birthplace	Residence
Outcome Mean	0.240	0.240	0.097	0.097	0.692	0.692
Percentage Effect	-1.9%	-1.8%	0.9%	-2.4%	-1.0%	-0.8%

Notes: Sample includes respondents from the 2005-2017 American Community Survey with a high school diploma or higher that were born in the U.S. and 22 years of age or older. Controls include binary variables for gender and race along with credit requirements in math, English, social studies, and science by high school graduation year and state of residence, controls for state level high school staffing, and availability of merit aid scholarships at the state level. Also included are state and high school graduation year fixed effects. Standard errors are clustered at the state level. *** p<0.01, ** p<0.05, * p<0.1

repayment progress after one year is correlated with long term repayment outcomes. Students who had made progress on their loans one year after entering repayment were half as likely to default and were 36% more likely to have paid off their full balance.

I find that median student loan balances are not significantly declining as a result of personal finance education mandates for first generation or low income students who are better at repaying loans. However, improvements in repayment rates for high income students might be a result of decreased borrowing. The results suggest the median high income student loan debt is 8% lower as a result of PFL mandates.

I use correct answers on financial literacy questions as a proxy for general financial literacy. I find no evidence that students bound by PFL mandates are more financial literate when surveyed. Across 13 questions asked in three surveys, I find no evidence that students bound by a personal finance mandate have a higher probability of correctly answering these questions. This does not necessarily imply that PFL mandates are ineffective at improving financial literacy. Rather, it is possible that improvements in financial literacy depreciate quickly after high school and/or non-mandated peers quickly catch up. In this case, personal finance education in high school may still operate as a just-in-time intervention in which financial literacy is temporarily improved at the same time postsecondary financing decisions are made.

On the other hand, I present evidence that students bound by personal finance mandates are more knowledgeable about the federal financial aid system. Students bound by mandates are more likely to correctly answer one of the three questions about federal

student loans and answer more of these questions correctly. This suggests that students bound by the mandates may be better able to repay student loans in part due to increased familiarity with the federal student loan system. If this is the case, personal finance mandates might not be necessary to improve student loan outcomes if the federal loan system were to be simplified. The results from this paper lend further evidence to the string of literature that shows potential benefits to a more streamlined federal financial aid system with fewer complexities that borrowers must learn before making postsecondary financing decisions (Dynarski and Scott-Clayton, 2006; Bettinger, Long, Oreopoulos, and Sanbonmatsu, 2012; Novak and McKinney, 2011; Dynarski and Scott-Clayton, 2013; Castleman, Schwartz, and Baum, 2015; Kofoed, 2017). However, the benefits to personal finance education mandates highlighted in this paper and in the related literature indeed suggest that mandating personal finance education in high school can improve financial outcomes for those students exposed to course material.

1.8 Appendix

1.8.1 Derivation of Motivating Specification

In this section, I show that when the three assumptions are satisfied, the aggregated estimating equation, Equation (1.2), consistently estimates the difference-in-differences parameter, γ^{DD}. First, assume the following assumptions hold:

1. Parallel Trends Assumption: $E[\Delta y_{0,ist} \mid D_{s(i)t} = 0] = E[\Delta y_{0,ist} \mid D_{s(i)t} = 1] \ \forall t$
2. Cohort Matching Assumption: $k_i = k \ \forall i$
3. Stability of University Mapping: $G(\ i, D_{s(i)t} = 1) = G(\ i, D_{s(i)t} = 0)$

where $s(i)$ is the state of high school for student i. First, define the function $G : \mathcal{I} \times \mathcal{D} \to \mathcal{J}$ where $\mathcal{I} = \{1, ..., I\}$, $\mathcal{D} = \{0, 1\}$, and $\mathcal{J} = \{0, 1, ..., J\}$. By Assumption 3, $G(i, D_{s(i)t} = 1) = G(i, D_{s(i)t} = 0)$ so we can simplify this function to G' which maps $\mathcal{I} \to \mathcal{J}$ such that $G'(i) = j$ is independent of $D_{s(i)t}$. Recall the difference-in-differences specification using micro-level data is:

$$y_{ist} = \alpha + \gamma^{DD} D_{s(i)t} + \delta_{s(i)t} + \varepsilon_{ist}, \tag{1.11}$$

Using the assumption that $k_i = k$ for all i, we can define $\tau := t + k$. Define $J_{j\tau}$ equal to $\{i : G'(i) = j, t = \tau - k\}$ and define $|J_{j\tau}|$ as the number of students in $J_{j\tau}$. The aggregated outcome, $Y_{j\tau}$, is defined by

$$Y_{j\tau} \coloneqq \frac{1}{|J_{j\tau}|} \sum_{i \in J_{j\tau}} y_{ist}$$

which constructs the average of y for all students in the set $J_{j\tau}$. Similarly, the same transformation can be applied to the RHS of Equation (1.11):

$$\begin{aligned}
Y_{j\tau} &= \frac{1}{|J_{j\tau}|} \sum_{i \in J_{j\tau}} \left[\alpha + \gamma^{DD} D_{s(i)t} + \delta_{s(i)t} + \varepsilon_{ist} \right] \\
&= \frac{1}{|J_{j\tau}|} \sum_{i \in J_{j\tau}} \alpha + \frac{1}{|J_{j\tau}|} \sum_{i \in J_{j\tau}} \left[\gamma^{DD} D_{s(i)t} \right] + \frac{1}{|J_{j\tau}|} \sum_{i \in J_{j\tau}} \delta_{s(i)t} + \frac{1}{|J_{j\tau}|} \sum_{i \in J_{j\tau}} \varepsilon_{ist} \\
&= \alpha + \gamma^{DD} \left[\frac{\sum\limits_{i \in J_{j\tau}} D_{s(i)t}}{|J_{j\tau}|} \right] + \frac{1}{|J_{j\tau}|} \sum_{i \in J_{j\tau}} \delta_{s(i)t} + \frac{1}{|J_{j\tau}|} \sum_{i \in J_{j\tau}} \varepsilon_{ist}
\end{aligned}$$

The first term in the specification trivially reduces to α. The second term reduces to the share of students in $J_{j\tau}$ for which $D_{s(i)t} = 1$ which we will define as $\text{pctBound}_{j\tau}$. Additionally, since the error term is assumed mean-zero in the micro-level case conditional on observables and the Parallel Trends Assumption, the aggregated university-level error draws will also be conditionally mean-zero since the allocation of students to universities is unchanged by $D_{s(i)t}$. As a result, the university error term can be rewritten as an arbitrary mean-zero error term $e_{j\tau}$.

$$Y_{j\tau} = \alpha + \gamma^{DD} \text{pctBound}_{j\tau} + \frac{1}{|J_{j\tau}|} \sum_{i \in J_{j\tau}} \delta_{s(i)t} + e_{j\tau}$$

By the Parallel Trends Assumption, we can rewrite $\delta_{s(i)t} = \delta_{s(i)} + \delta_t$. Further, $\delta_{s(i)}$ can be rewritten as $\sum\limits_{s=1}^{S} \delta_s \cdot 1\{s(i) = s\}$ and the specification becomes

$$Y_{j\tau} = \alpha + \gamma^{DD} \text{pctBound}_{j\tau} + \sum_{i \in J_{j\tau}} \sum_{s=1}^{S} \delta_s \frac{1 \cdot \left\{ s(i) = s, i \in J_{j\tau} \right\}}{|J_{j\tau}|} + \frac{1}{|J_{j\tau}|} \sum_{i \in J_{j\tau}} \delta_t + e_{j\tau}$$

$$Y_{j\tau} = \alpha + \gamma^{DD} \text{pctBound}_{j\tau} + \delta_t + \sum_{s=1}^{S} \delta_s \left[\sum_{i \in J_{j\tau}} \frac{1 \cdot \left\{ s(i) = s, i \in J_{j\tau} \right\}}{|J_{j\tau}|} \right] + e_{j\tau}$$

Since $\tau = t + k$ by assumption, the time fixed effect is unchanged and δ_τ is just a change in notation. However, the last remaining term is more nuanced. Note that this term is a reweighting of the feeder-state fixed effect in accordance with the share of the cohort from

each feeder state. For ease of interpretation, define the following terms

$$\text{StateShare}_{sj\tau} \coloneqq \sum_{i \in J_{j\tau}} \frac{1 \cdot \{s(i) = s, i \in J_{j\tau}\}}{|J_{j\tau}|}, \quad \delta_j \coloneqq \sum_{\tau=k}^{T+k} \sum_{s=1}^{S} \delta_s \text{StateShare}_{sj\tau}$$

Adding and subtracting δ_j yields:

$$\begin{aligned} Y_{j\tau} &= \alpha + \gamma^{DD} \text{pctBound}_{j\tau} + \delta_t + \delta_j - \sum_{\tau=k}^{T+k} \sum_{s=1}^{S} \delta_s \text{StateShare}_{sj\tau} + \sum_{s=1}^{S} \delta_s \text{StateShare}_{sj\tau} + e_{j\tau} \\ &= \alpha + \gamma^{DD} \text{pctBound}_{j\tau} + \delta_t + \delta_j + \left[\sum_{s=1}^{S} \delta_s \left(\text{StateShare}_{sj\tau} - \sum_{\tau=k}^{T+k} \text{StateShare}_{sj\tau} \right) \right] + e_{j\tau} \end{aligned}$$

The remaining term in brackets represents the sum of transitory deviations from the university's mean share of students from each state multiplied by the fixed effect for each state. By the Stability of University Mapping assumption, this term is independent of the components of $\text{pctBound}_{j\tau}$. Collecting this transitory enrollment deviations term with $e_{j\tau}$, we can rewrite the estimating equation as:

$$Y_{j\tau} = \alpha + \gamma^{DD} \text{pctBound}_{j\tau} + \delta_j + \delta_t + v_{j\tau}. \tag{1.12}$$

Hence, under the three aforementioned assumptions, the aggregate university-level specification consistently estimates the micro-level difference-in-differences specification. Additionally, in Section 1.6.2, I estimate an alternative university-level specification that holds $\text{StateShare}_{sj\tau}$ fixed at initial levels. In this specification, the identification of γ^{DD} in the university-level specification is consistent *even in the case where students alter their college choice* as a result of $D_{s(i)t}$.

1.8.2 Data Appendix

College Scorecard

All data used in the analysis was pulled from the College Scorecard website using the October 30, 2018 update. The subsequent updates (as of September 27, 2019) did not affect the measures from the NSLDS used in this paper .

Integrated Postsecondary Education Data System

I collect previous state of residence data from the Integrated Postsecondary Education Data System (IPEDS). This data comes from the Compare Institutions tool on the IPEDS website. The data include years 1986 through 2016 for all universities in the Scorecard sample. From the Fall Enrollment category, I use the "State of residence when student was first admitted" and counts of "First-time degree/certificate-seeking undergraduate students who graduated from high school in the past 12 months." These data were required to be submitted in: 2016, 2014, 2012, 2010, 2008, 2006, 2004, 2002, 2000, 1998, 1996, 1994, 1992, 1988, and 1986. Universities could voluntarily provide this data in: 2015, 2013, 2011, 2009, 2007, 2005, 2003, and 2001. I impute missing values by linearly interpolating between the nearest non-missing years. In addition, Section 1.8.4 estimates Equation (1.4) by using instrumented values of enrollment counts rather than linear interpolation.

High School Staffing Variables

I collect counts of state level high school staffing to use as controls in all specifications. These data come from the Common Core of Data (CCD) and are accessed using the `educationdata` Stata package from the Urban Institute. I pull these data for the years 1993 through 2015 at the school district level. Counts for each of the following are collected and aggregated to the state level: total staff, full-time equivalent total teachers, full-time equivalent total school support staff, total school guidance counselors, and total student support staff.

High School Graduation Requirements

I create a panel dataset of credit requirements for high school graduation at the state-by-graduation-year level. These data are primarily sourced from the National Center for Education Statistics (NCES) Digest of Education Statistics Chapter 2. These tables present snapshots in time of state credit requirements for each state along with the first effective graduating cohort bound by the requirements. The first table is from 1995 and I use these snapshots to track changes in graduation requirements in: **Total Credits, English/Language Arts, Social Studies, Math, and Science**. The creation of this data required some decisions in which I try to follow objective rules. First, not all states have state requirements for high school graduation. States like Colorado deferred requirements to the district level. For these states, I impute the state requirements by substituting the national average for each graduating cohort for states with requirements and I include a binary variable denoting local control. Second, many states have multiple tracts students

can select with different credit requirements for each track. When possible, I select the vector of graduating requirements that had the minimum standards. These are typically obvious when the choice is between a "standard" diploma and an "honors" diploma, however the definition can be more subjective when states allow students a technical career path. In these cases, I choose the standard diploma requirements as the technical career path students are less likely to attend a four-year college after high school graduation.

I supplement and cross reference the NCES data with data from the Education Commission of the States 50-State Comparison: High School Graduation Requirements (Macdonald, Dounay-Zinth, and Pompelia, 2019). When conflicts between the sources arose, I tracked the course standards using state Department of Education websites to resolve discrepancies. This data is available upon request.

State Merit Aid

I use the definition of state merit aid availability at the state-graduation-year level as defined by Sjoquist and Winters (2015). They define merit aid scholarships as "strong" and "weak" merit aid programs and I follow their convention. I include a binary indicator variable at the high school graduating cohort by year level for the presence of weak and strong merit aid in each specification.

Constructing University Cohort Controls from State-by-Graduation-Year Data

A vector of incoming cohort level controls are included in $\mathbf{X}_{it}$. In a similar manner to Equation (1.3), I create a vector of control variables for each incoming university cohort that is weighted by the state composition of the incoming cohort. I use high school graduation state j by high school graduation year t variables, x_{jt}, combined with previous state of residence data, enroll_{ijt}, for university i from state j in year t to construct an incoming university cohort measure for each variable in X_{it}:

$$X_{it} = \frac{\sum_{j=1}^{51} x_{jt} \times \text{enroll}_{ijt}}{\sum_{j=1}^{51} \text{enroll}_{ijt}}, \tag{1.13}$$

This vector includes the state level measures of high school staffing and high school graduation requirements.[42] In addition to these state weighted controls, $\mathbf{X}_{it}$ also includes

[42]Not all states have high school graduation standards set at the state level. For states with no state standards, the mean value across all states is used and a binary variable is included denoting local control of high school graduation standards.

binary variables for whether the state of university offered a merit aid scholarship along with unemployment rates for periods t through $t + k$.

1.8.3 Randomization Inference Algorithm

The randomization inference algorithm used to compute the empirical p-values is based off the RI-β algorithm in MacKinnon and Webb (forthcoming). I conduct 3000 replications of Equation (1.4) for each outcome variable where the identifying variation in the replication is randomly generated by supposing that the adopting states do not adopt and the non-adopting states do adopt. In each of these replications, it should be the case that the estimated treatment effect for the placebo replications is zero on average. Further, the estimated treatment effect using the observed pctBound measure should be a sufficiently extreme value in the distribution of placebo replications. The algorithm proceeds as follows for each replication:

1. Split the sample of 50 states plus D.C. into two groups

   Group A: States adopting a mandate binding for the class of 2008 and prior (13 states)

   Group B: States adopting a mandate binding for the class of 2009 and later and states that never adopt a mandate.

2. Choose 13 states at random from Group B to slot into the mandate adoption slots observed in the true data[43]

3. Use this selection of states and adoption years to generate placebo pfMandate_{jt}.

4. Compute $\text{pctBound}_{it} = \frac{\sum_{j=1}^{51} \text{pfMandate}_{jt} \times \text{enroll}_{ijt}}{\sum_{j=1}^{51} \text{enroll}_{ijt}}$ using placebo pfMandate_{jt}.

5. Estimate Equation (1.4) using the placebo pctBound_{it}.

6. Store $\hat{\gamma}_n$.

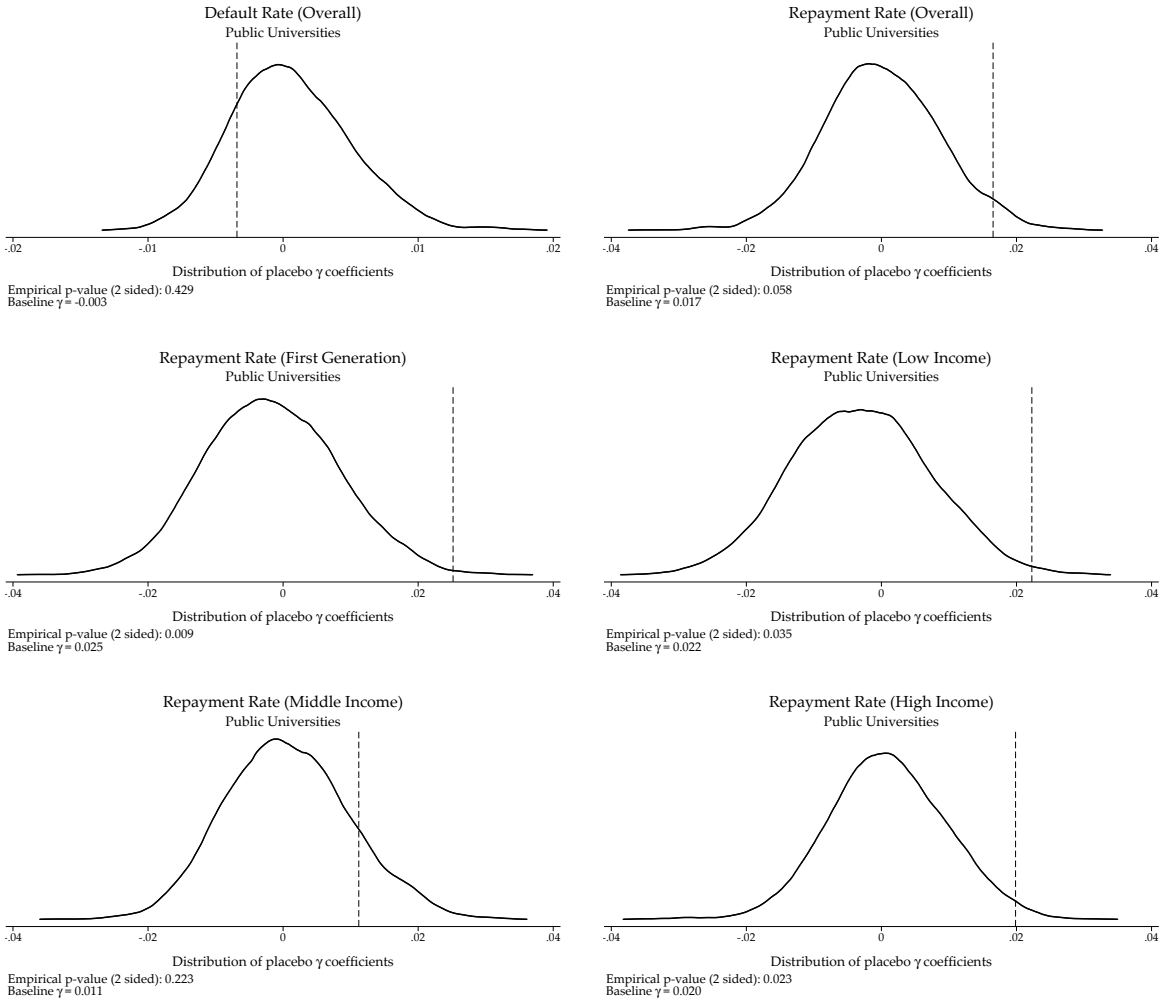
Once all $\hat{\gamma}_n$ for $n = 1, \ldots, 3000$ are collected, the empirical p-value is computed using:

$$\bar{p} = \frac{1}{3000} \sum_{n=1}^{3000} 1 \cdot \left\{ |\hat{\gamma}_n| \geq |\hat{\gamma}_{true}| \right\} \tag{1.14}$$

[43] One adopting state in 1993, 1996, 1998, 2002, 2006 and 2008. Three adopting states in 2005. Four

Figure 1.11 presents the distributions of the $\hat{\gamma}_n$ estimates for each outcome for the public university sample with the $\hat{\gamma}$ from the baseline specification marked in each distribution.

Figure 1.11: Distribution of Placebo Estimates for pctBound for Repayment Outcomes for Public Universities



1.8.4 Alternative Specifications

In this section, I explore whether the results presented above are sensitive to the choice of specification and the use of the continuous treatment measure.

adopting states in 2007.

State-Level Difference-in-Differences for High In-state Universities

First, I abandon the use of the continuous treatment measure to estimate a more straight-forward difference-in-differences specification in which treatment status is assigned to each university at the state level. To isolate the sample to only those that are most affected by the within-state adoption of a personal finance education mandate, I restrict the sample to public and private universities with a historically high in-state percentage of students. For each university, I calculate the mean percentage of students who resided in the state in the previous year over the sample years and only include a university in this analysis if the mean percentage of in-state students is 70% or higher.[44] The result is a sample of 656 universities of which 370 are public and 286 are private. This subsample represents over 75% of the public universities in the sample but less than one-third of the private universities. In this specification, each university is assumed to only be affected by its own state's mandate adoption (if any) and universities in states that do not adopt a mandate act as controls. The specification is similar to Equation (1.4)

$$y_{is,t+6} = \gamma \text{pfMandate}_{st} + \beta \mathbf{X}_{st} + \delta_s + \delta_t + \varepsilon_{ist}, \tag{1.15}$$

where $y_{is,t+6}$ is the same student loan repayment outcome for university i located in state s for the repayment cohort matched to high school graduating class t. Rather than pctBound_{it} as in Equation (1.4), pfMandate_{st} is equal to one if the university state s has a binding mandate for high school graduating cohort t. Also included are the vector of control variables $\mathbf{X}_{st}$ at the state level which include other course credit requirements, high school staffing levels, availability of state merit aid scholarships, and a vector of the state unemployment rates between periods t and $t+6$. Fixed effects for state (δ_s) and high school graduating cohorts (δ_t) are also included and standard errors are clustered at the state level.

Table 1.10 reports the estimated γ coefficients for this specification for all universities with 70% or higher historical in-state percentage for the main outcome variable split by public and private universities. Columns 3 and 4 in Panel A show improvements in the one-year repayment rate for first generation and low income students similar to those found in Table 1.4. However, the estimates presented in Panel B suggest that private universities who receive a large share of students from in-state high schools indeed see improvements in student loan repayment at least for low income students. In fact, the point estimates for private universities are larger than for public universities. This divergence in

[44] For the sample of universities, the median historical in-state percentage is 64% so this is roughly half the universities in the main analysis

the sample of private universities across specifications may be due to the type of student attending out-of-state private universities or the type of private universities that attract largely in-state students.

Instrumenting for Enrollment Counts

As noted above, universities are required to send data on the previous state of residence for each incoming cohort only in even numbered years. Universities may elect to also provide this information in odd years but are not required. As a result, the data contain many missing values over the sample. Further, investigation of the data reveal numerous transcription errors in which the cohort is coded as including only students who graduated from high school longer than 12 months prior when this is highly unlikely given previous years' data. In the main analysis, I linearly interpolate missing values using the neighboring non-missing years. However, trends in attendance may not be linear in years and idiosyncratic and transitory shocks to attendance numbers may occur which deviate from linearly interpolated values.

In this section, I conduct a more thorough exercise to replace missing data values that uses more information to predict missing values by instrumenting enroll_{isjt} with linear and quadratic trends, a series of fixed effects, and the availability of state merit aid scholarships. Equation (1.16) details the specification for this strategy:

$$\begin{aligned}\text{enroll}_{isjt} =& \delta_i + t \cdot \delta_i + t^2 \cdot \delta_i + t \cdot \delta_s + t^2 \cdot \delta_s \\ &+ t \cdot \delta_{ij} + t^2 \cdot \delta_{ij} + \text{meritAid}_{jt} + \text{meritAid}_{jt} \cdot \{s = j\} \\ &+ \delta_{jt} + \varepsilon_{isjt}\end{aligned} \tag{1.16}$$

In this specification, predicted values of enroll_{isjt} are estimated by regressing enroll_{isjt} on university fixed effects and state fixed effects both of which are interacted with linear and quadratic time trends. In addition, linear and quadratic trends for each university-by-feeder state are also included. I include an indicator for whether the feeder state offered a state merit aid scholarship for cohort t. Since state merit aid scholarships provide an added incentive to attend an in-state school, the addition of a scholarship might cause students to be less likely to attend an out-of-state school (Fitzpatrick and Jones, 2016). For this reason, I also include an interaction of meritAid_{jt} with an indicator for whether the feeder state is an in-state university since this effect would be opposite-signed. Lastly, I include feeder-state-specific year fixed effects to capture transitory shocks to feeder state level college enrollment.

I use the estimated coefficients and fixed effects to predict enroll_{isjt} ($\widetilde{\text{enroll}}_{isjt}$) for both

non-missing values included in the regression as well as missing observations not included in the regression. I then use $\widetilde{\text{enroll}}_{isjt}$ to construct $\widetilde{\text{pctBound}}_{ist}$ as in Equation (1.3) to re-estimate Equation (1.4). Table 1.11 presents the results from this exercise. The estimates in Panel A for public universities are largely consistent with the estimates presented in Table 1.4. However, the results in Panel B potentially suggest better outcomes for private university students as a result of changes in $\widetilde{\text{enroll}}_{isjt}$. This result suggests that private universities may be more likely to choose not to report enrollment data in odd years resulting in more missing data and thus less precise results during linear interpolation in the main results. While the point estimates are larger and suggest improvements in the repayment rate for first generation and low income students, these estimates are not statistically significant at conventional levels. Further, this estimation does not take into account the fact that $\widetilde{\text{enroll}}_{isjt}$ is a generated regressor and thus standard errors are likely under-estimated as is.

Table 1.10: Robustness: Difference-in-Differences for High In-state University Sample for Public and Private Universities

	Default Rate	Repayment Rate				
	(1)	(2)	(3)	(4)	(5)	(6)
A. Public	Overall	Overall	First Gen	Low Income	Middle Income	High Income
pfMandate	-0.002	0.015	0.023**	0.021**	0.009	0.017
	(0.002)	(0.010)	(0.011)	(0.009)	(0.011)	(0.011)
Universities	370	370	370	370	366	366
Cohorts	1993-2006	2001-2008	2001-2008	2001-2008	2001-2008	2001-2008
Outcome Mean	0.046	0.589	0.552	0.479	0.628	0.723
Percentage Effect	-3.9%	2.6%	4.2%	4.4%	1.5%	2.4%
B. Private						
pfMandate	-0.005	0.016	0.027	0.031*	0.014	0.002
	(0.003)	(0.017)	(0.018)	(0.018)	(0.016)	(0.016)
Universities	286	286	284	285	271	268
Cohorts	1993-2006	2001-2008	2001-2008	2001-2008	2001-2008	2001-2008
Outcome Mean	0.040	0.571	0.522	0.421	0.618	0.747
Percentage Effect	-11.5%	2.8%	5.1%	7.3%	2.2%	0.2%

Regressions are weighted using the number of students used to compute each outcome metric. Each column reports a coefficient from a separate regression where the independent variable is pfMandate_{jt} and the outcome is denoted in the column header. The sample includes public and private four-year universities with 70% or higher in-state share of students during the sample period. Default rate analysis includes high school graduating classes 1993 through 2006 and repayment rate analysis includes high school graduating classes 2001 through 2008 due to data availability. First Gen students are defined as students whose parents did not have a college degree. Low Income, Middle Income, and High Income students are defined as household income less than 30,000, between 30,000 and 75,000 and above 75,000, respectively. Controls include cohort weighted credit requirements in math, English, social studies, and science by high school graduation cohort and controls for state level high school staffing, and availability of merit aid scholarships. Also included are university and high school graduation year fixed effects. Standard errors clustered at the state level are presented in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

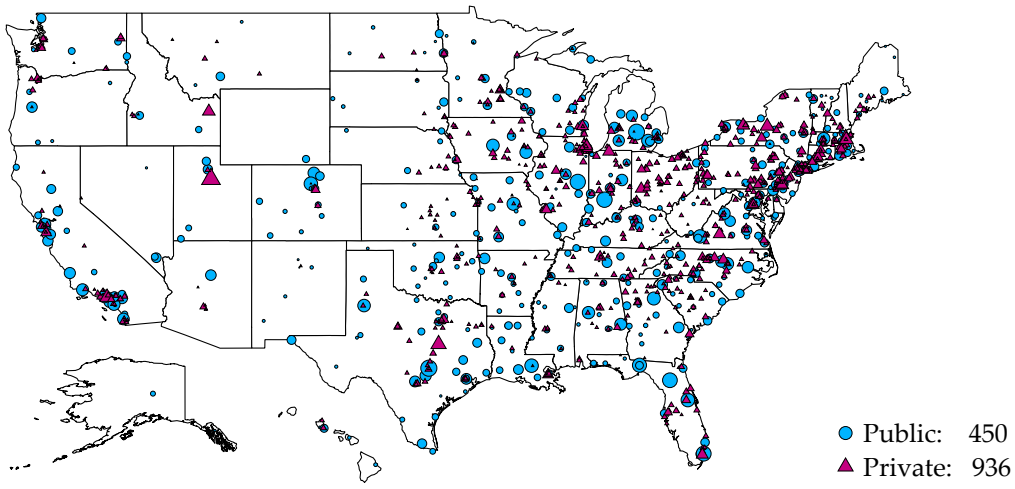
Table 1.11: Robustness: Dose Response Estimates with Instrumented Enrollment for Public and Private Universities

	Default Rate	Repayment Rate				
	(1)	(2)	(3)	(4)	(5)	(6)
A. Public	Overall	Overall	First Gen	Low Income	Middle Income	High Income
pctBound	-0.003	0.016	0.025**	0.022**	0.010	0.018
	(0.002)	(0.009)	(0.010)	(0.010)	(0.010)	(0.011)
Universities	444	446	446	445	442	442
Cohorts	1993-2006	2001-2008	2001-2008	2001-2008	2001-2008	2001-2008
Outcome Mean	0.046	0.594	0.551	0.478	0.631	0.730
Percentage Effect	-5.6%	2.6%	4.5%	4.6%	1.6%	2.5%
B. Private						
pctBound	0.002	-0.000	0.012	0.017	-0.002	-0.009
	(0.005)	(0.014)	(0.011)	(0.017)	(0.009)	(0.010)
Universities	823	821	800	804	796	799
Cohorts	1993-2006	2001-2008	2001-2008	2001-2008	2001-2008	2001-2008
Outcome Mean	0.035	0.600	0.516	0.426	0.636	0.779
Percentage Effect	6.2%	-0.1%	2.4%	4.1%	-0.4%	-1.2%

Regressions are weighted using the number of students used to compute each outcome metric. Each column reports a coefficient from a separate regression where the independent variable is pctBound and the outcome is denoted in the column header. The sample includes public and private four-year universities. Default rate analysis includes high school graduating classes 1993 through 2006 and repayment rate analysis includes high school graduating classes 2001 through 2008 due to data availability. First Gen students are defined as students whose parents did not have a college degree. Low Income, Middle Income, and High Income students are defined as household income less than 30,000, between 30,000 and 75,000 and above 75,000, respectively. Controls include cohort weighted credit requirements in math, English, social studies, and science by high school graduation cohort and controls for state level high school staffing, and availability of merit aid scholarships. Also included are university and high school graduation year fixed effects. Standard errors clustered at the state level are presented in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

1.8.5 Additional Tables and Figures

Figure 1.12: Map of University Sample from College Scorecard



The map above shows the locations and relative cohort sizes of the sample of public and private universities from the college scorecard. Each marker is weighted by the mean cohort size over the sample period. Public universities are denoted with a blue circle while private universities are denoted by a pink triangle. Some universities in the sample have missing GPS coordinates and are not plotted despite inclusion in the sample.

Table 1.12: Sample Construction

Restriction	All Universities	Public Universities	Private Universities
Full Sample	3,563	708	2,855
Balanced Sample	1,844	590	1,254
Single Branch	1,498	470	1,028
Non-missing Outcomes	1,386	450	936

The table above describes the number of universities that survive each iterative step of creating the sample. The first row is the full sample of four-year universities from the College Scorecard database. The second row is the result of removing universities that opened or closed during the sample period. The third row is the result of removing universities with outcome data aggregated across multiple branches. The fourth row is the result of removing universities with missing outcome data due to non-Title IV status or all cell sizes smaller than 30 students and thus suppressed.

Figure 1.13: Word Cloud of PFL State Standard Text

The image above shows a word cloud of the text for the state PFL standards for the following states: Arizona, Arkansas, Colorado, Florida, Georgia, Idaho, Iowa, Kansas, Louisana, Michigan, Missouri, New Hampshire, New Jersey, New York, North Carolina, Oregon, South Carolina, South Dakota, Tennessee, Texas, Utah, and Virginia.

Table 1.13: Events per Academic Year

Academic Year	Events	Adopting States
1996	89	New York
1997	0	
1998	34	Michigan
1999	0	
2000	0	
2001	0	
2002	1	Wyoming
2003	0	
2004	0	
2005	41	Arizona, Arkansas, Louisiana
2006	10	South Dakota
2007	130	Georgia, Idaho, North Carolina, Texas
2008	6	Utah
2009	38	Colorado, South Carolina
2010	29	Missouri
2011	68	Iowa, New Jersey, Tennessee
2012	12	Kansas
2013	13	Oregon
2014	53	Florida, Virginia
Total	524	

The table above details the number of university events in each academic year where an event is defined as a year-over-year change in pctBound_{ist} of 50 percentage points or larger. In addition, the last column summarizes the states that adopt a personal finance mandate in each academic year. Events induced by New Hampshire's 1993 mandate occur before the sample period for outcome data.

Table 1.14: Difference-in-Differences Estimates for Changes to College-going for Two Year Universities (2001-2008)

	(1) pctAny	(2) pctPublic	(3) pctNonProfit	(4) pctForProfit	(5) pctInstate
PF Mandate	0.005	0.006	-0.000	-0.000	0.004
	(0.006)	(0.006)	(0.000)	(0.001)	(0.006)
Observations	408	408	408	408	408
Cohorts	2001-2008	2001-2008	2001-2008	2001-2008	2001-2008
Outcome Mean	0.161	0.152	0.002	0.007	0.146
Percentage Effect	3.3%	3.9%	-5.0%	-2.9%	2.9%

Each column above reports the Difference-in-Differences estimate for the outcome in each column header. Each outcome measures the percentage of recent high school graduates from a state who chose to attend a two year university of the given charactistic. Each observation is a state-year cell. The sample is restricted to high school graduation years 2001 through 2008 to match the results for the one year repayment rates. pctPublic and pctPrivate sum to one and thus the results are inverses of each other. Controls include credit requirements in math, English, social studies, and science by high school graduation year and state of university, controls for state level high school staffing, and availability of merit aid scholarships at the state level. Also included are state and high school graduation year fixed effects. *** p<0.01, ** p<0.05, * p<0.1

Table 1.15: Difference-in-Differences Estimates for Changes to College-going for Four Year Universities (2001-2008)

	(1) pctAny	(2) pctPublic	(3) pctNonProfit	(4) pctForProfit	(5) pctInstate
PF Mandate	-0.007	-0.003	-0.004	-0.000	-0.004
	(0.007)	(0.005)	(0.003)	(0.001)	(0.007)
Observations	408	408	408	408	408
Cohorts	2001-2008	2001-2008	2001-2008	2001-2008	2001-2008
Outcome Mean	0.439	0.296	0.137	0.006	0.292
Percentage Effect	-1.7%	-1.1%	-2.9%	-2.3%	-1.3%

Each column above reports the Difference-in-Differences estimate for the outcome in each column header. Each outcome measures the percentage of recent high school graduates from a state who chose to attend a four year university of the given charactistic. Each observation is a state-year cell. The sample is restricted to high school graduation years 2001 through 2008 to match the results for the one year repayment rates. pctPublic and pctPrivate sum to one and thus the results are inverses of each other. Controls include credit requirements in math, English, social studies, and science by high school graduation year and state of university, controls for state level high school staffing, and availability of merit aid scholarships at the state level. Also included are state and high school graduation year fixed effects. *** p<0.01, ** p<0.05, * p<0.1

Table 1.16: Robustness: Stable State Composition Estimates for Public and Private Universities

	Default Rate	Repayment Rate				
	(1)	(2)	(3)	(4)	(5)	(6)
A. Public	Overall	Overall	First Gen	Low Income	Middle Income	High Income
$\widehat{\text{pctBound}}$	-0.003	0.018*	0.028**	0.026**	0.012	0.020*
	(0.002)	(0.010)	(0.011)	(0.010)	(0.011)	(0.012)
Universities	450	450	449	449	445	445
Cohorts	1993-2006	2001-2008	2001-2008	2001-2008	2001-2008	2001-2008
Outcome Mean	0.046	0.593	0.551	0.478	0.630	0.730
Percentage Effect	-5.8%	3.0%	5.0%	5.3%	1.8%	2.7%
B. Private						
$\widehat{\text{pctBound}}$	-0.003	0.003	0.025**	0.024	0.002	-0.011
	(0.004)	(0.017)	(0.012)	(0.020)	(0.012)	(0.013)
Universities	936	934	891	905	875	873
Cohorts	1993-2006	2001-2008	2001-2008	2001-2008	2001-2008	2001-2008
Outcome Mean	0.036	0.595	0.512	0.423	0.632	0.777
Percentage Effect	-9.3%	0.6%	5.0%	5.6%	0.3%	-1.4%

Regressions are weighted using the number of students used to compute each outcome metric. Each column reports a coefficient from a separate regression where the independent variable is $\widehat{\text{pctBound}}$ and the outcome is denoted in the column header. The sample includes public and private four-year universities. Default rate analysis includes high school graduating classes 1993 through 2006 and repayment rate analysis includes high school graduating classes 2001 through 2008 due to data availability. First Gen students are defined as students whose parents did not have a college degree. Low Income, Middle Income, and High Income students are defined as household income less than 30,000, between 30,000 and 75,000 and above 75,000, respectively. Controls include cohort weighted credit requirements in math, English, social studies, and science by high school graduation cohort and controls for state level high school staffing, and availability of merit aid scholarships. Also included are university and high school graduation year fixed effects. Standard errors clustered at the state level are presented in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

Table 1.17: Difference-in-Differences Estimates for Financial Literacy and Loan Literacy from NPSAS

	(1) FL1: Interest	(2) FL2: Inflation	(3) FL3: Risk	(4) FL: Num Correct	(5) LL1: Credit	(6) LL2: Garnish Wages	(7) LL3: Tax Returns	(8) LL: Num Correct
pfMandate	0.004	-0.002	0.002	0.004	0.013	0.033**	0.018	0.064**
	(0.020)	(0.011)	(0.019)	(0.040)	(0.015)	(0.015)	(0.016)	(0.030)
Observations	45,230	45,230	45,230	45,230	45,230	45,230	45,230	45,230
Cohorts	2009-2016	2009-2016	2009-2016	2009-2016	2009-2016	2009-2016	2009-2016	2009-2016
Outcome Mean	0.662	0.856	0.466	1.944	0.753	0.558	0.660	1.97
Percentage Effect	0.5%	-0.4%	0.2%	0.1%	1.7%	6.0%	2.7%	3.2%

Source: U.S Department of Education, National Center for Education Statistics, Restricted-use National Postsecondary Student Aid Study 2016. Each column reports the coefficient of the binary personal finance mandate variable from a separate linear regression. Columns 1-3 and 5-7 report the impact of a personal finance mandate on an indicator variable for whether the respondent correctly answered the question from a linear probability model. Columns 4 and 8 show the impact of the mandate on the total number of correct questions answered. Each regression includes controls for race, gender, Expected Family Contribution, year of schooling, and public or private high school attended. Also included are state of high school attendance and high school graduation year fixed effects. Standard errors are clustered at the state level. *** p<0.01, ** p<0.05, * p<0.1

Table 1.18: Difference-in-Difference Estimates for Financial Literacy from NFCS

	(1) Q1: Interest	(2) Q2: Inflation	(3) Q3: Risk	(4) Q4: Bonds	(5) Q5: Mortgage	(6) Number Correct
pfMandate	0.027	0.042	0.009	-0.038	0.020	0.060
	(0.031)	(0.030)	(0.027)	(0.024)	(0.030)	(0.110)
Observations	10,020	10,020	10,020	10,020	10,020	10,020
Cohorts	2001-2008	2001-2008	2001-2008	2001-2008	2001-2008	2001-2008
Outcome Mean	0.714	0.420	0.363	0.197	0.711	2.405
Percentage Effect	3.9%	10.0%	2.4%	-19.3%	2.8%	2.5%

Notes: Sample includes respondents from the 2012, 2015, and 2018 waves of the restricted use National Financial Capabilities Survey with a high school diploma or higher. Controls include binary variables for gender, race, and education along with credit requirements in math, English, social studies, and science by high school graduation year and state of residence, controls for state level high school staffing, and availability of merit aid scholarships at the state level. Also included are state and high school graduation year fixed effects and state-level survey weights. Standard errors are clustered at the state level. *** p<0.01, ** p<0.05, * p<0.1

Table 1.19: Difference-in-Differences Estimates for Financial Literacy from the Survey of Household Economic Decision-making

	(1) Q1: Interest	(2) Q2: Inflation	(3) Q3: Risk	(4) Q4: Returns	(5) Q5: Housing	(6) Number Correct
pfMandate	0.002	-0.051	-0.014	0.015	-0.094	-0.143
	(0.050)	(0.039)	(0.051)	(0.048)	(0.058)	(0.109)
Observations	2,604	2,604	2,604	2,604	2,604	2,604
Cohorts	2001-2008	2001-2008	2001-2008	2001-2008	2001-2008	2001-2008
Outcome Mean	0.653	0.478	0.425	0.367	0.579	2.502
Percentage Effect	0.3%	-10.7%	-3.2%	4.0%	-16.3%	-5.7%

Notes: Sample includes respondents from the 2017 and 2018 waves of the Survey of Household Economic Decision-making with a high school diploma or higher that graduated high school between 2001 and 2008. Controls include binary variables for gender, race, and education along with credit requirements in math, English, social studies, and science by high school graduation year and state of residence, controls for state level high school staffing, and availability of merit aid scholarships at the state level. Also included are state, survey wave, and high school graduation year fixed effects. Standard errors are clustered at the state level. *** p<0.01, ** p<0.05, * p<0.1

Table 1.20: Question Text for Financial Literacy Questions

A. NPSAS:16

Label	Question Text	Choices
FL1: Interest	Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, how much would you be able to buy with the money in this account?	More than today Exactly the same **Less than today**
FL2: Inflation	Suppose you had $100 in a savings account and the interest was 2% per year. After 5 years, how much do you think you would have in the account if you left the money to grow?	**More than $102** Exactly $102 Less than $102
FL3: Risk	Buying a single company's stock usually provides a safer return than a stock mutual fund.	True **False**
LL1: Credit	If a borrower is unable to repay his or her federal student loan, the government can report that the student debt is past due to the credit bureaus	**True** False
LL2: Garnish Wages	If a borrower is unable to repay his or her federal student loan, the government can have the student's employer withhold money from his or her pay (garnish wages) until the debt, plus any interest and fees, is repaid	**True** False
LL3: Tax Returns	If a borrower is unable to repay his or her federal student loan, the government can retain tax refunds and Social Security payments until the debt, plus any interest and fees, is repaid	**True** False
B. NFCS		
Q1: Interest	Suppose you had $100 in a savings account and the interest rate was 2% per year. After 5 years, how much do you think you would have in the account if you left the money to grow?	**More than $102** Exactly $102 Less than $102
Q2: Inflation	Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, how much would you be able to buy with the money in this account?	More than today Exactly the same **Less than today**
Q3: Risk	Buying a single company's stock usually provides a safer return than a stock mutual fund.	True **False**
Q4: Bonds	If interest rates rise, what will typically happen to bond prices?	They will rise **They will fall** They will stay the same No relationship
Q5: Mortgage	A 15-year mortgage typically requires higher monthly payments than a 30-year mortgage, but the total interest paid over the life of the loan will be less.	**True** False
C. SHED		
Q1: Interest	Suppose you had $100 in a savings account and the interest rate was 2% per year. After 5 years, how much do you think you would have in the account if you left the money to grow?	**More than $102** Exactly $102 Less than $102
Q2: Inflation	Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, how much would you be able to buy with the money in this account?	More than today Exactly the same **Less than today**
Q3: Risk	Buying a single company's stock usually provides a safer return than a stock mutual fund.	True **False**
Q4: Returns	Considering a long time period (for example 10 or 20 years), which asset described below normally gives the highest returns?	**Stocks** Bonds Savings accounts Precious metals
Q5: Housing	Housing prices in the US can never go down.	True **False**

Chapter 2: You're Not You When You're Hungry: Measuring The Impact of a Supplemental Nutrition Program on Childhood Test Scores

2.1 Introduction

Food insecurity in the United States remains a topic of public concern despite federal, state, and local programs designed to alleviate hunger for low income households. In 2018, almost 14% of households with children in the United States experienced at least one instance of food insecurity during the year. Incidence of food insecurity varies across region with the lowest rates in the Midwest and the highest in the South. Mississippi has had the highest rate of food insecurity since 2010 with one in eight households experiencing bouts of nutritional scarcity. Recent research finds that lack of nutrition can cause children to lose focus and perform worse in school, however interventions that relieve food insecurity are shown to improve test scores and reduce behavioral issues.(Frisvold, 2015; Schwartz and Rothbart, 2017; Figlio and Winicki, 2005; Maluccio et al., 2009).

In this paper, I explore the impact of a supplemental nutrition program targeting low socioeconomic status children in the Mississippi Delta region on standardized test scores and attendance patterns. The intervention was designed to replicate the Free and Reduced Lunch Program (FRLP) students receive during the week to last over the course of the weekend for children in grades three through five at two schools in the Mississippi Delta region. A survey provided to teachers and parents suggests the intervention increased attendance on Fridays and resulted in fewer behavioral issues. This paper more rigorously studies the effect of the intervention by employing panel data on grade-by-school test scores to estimate the effect of the supplemental nutrition program on mean test scores, the percentage of students achieving at different thresholds, and daily attendance. I use a difference-in-differences design to compare the schools selected for treatment to a set of schools who were unaffected. I find that the one-time intervention improved students' test scores, particularly for language arts. The improvements in mean scores are largely driven by shifts away from students achieving the lowest achievement threshold and toward achievement of a proficient standard.

I also use a triple-difference design where I incorporate administrative daily attendance records and incorporate kindergarten through second grade students who were not included in the intervention as another dimension for comparison. I find that attendance was higher for students selected for treatment and the improvements were concentrated on Fridays, Mondays, and Tuesdays. The improvement in Friday attendance is likely due to the transfer effect of receiving the bundles of food. However, improvements in Monday

and Tuesday attendance suggest that students had better nutrition over the course of the weekend.

The remainder of the paper is organized as follows: Section 2.2 reviews the literature for various nutrition interventions across the world to serve as benchmark treatment effects. Section 2.3 discusses the institutional details of this intervention. Section 2.4 describes the data used in this study. Section 2.5 discusses the empirical strategy employed to estimate the casual impact of the intervention. Section 2.6 summarizes the results of the analysis and Section 2.7 concludes the paper.

2.2 Literature

The link between nutrition and education has been explored extensively in both developed and developing countries. Many early studies were largely observational or relied on targeting children exhibiting signs of malnutrition. A study of 3,055 third-grade students in Vietnam examined the relationship between anthropometric status and educational achievement and found low test scores in mathematics and Vietnamese were correlated with both low height for age and low weight for age after controlling for age, sex, and school (Hall et al., 2001). A study in Chile surveyed a random sample of children graduating elementary school and high school and the results suggested academic achievement is positively correlated with consumption of dairy, meat, and eggs while consumption of fruits and vegetables is negatively correlated with low academic achievement. Food habits explained nearly 24% of variation in achievement for elementary school students (Ivanovic et al., 1992). A study in the United States used data from the Third National Health and Nutritional Examination Survey to test correlations between food insufficiency and academic achievement and other behavioral outcomes for respondents answering positively to instances of food insufficiency. After controlling for various covariates, the study finds students between 6 and 11 years old were more likely to report not getting along with other students, more likely to repeat a grade, have poorer arithmetic scores, and more likely to have seen a psychologist. Separately studying teenage respondents revealed similar results for behavior responses but no significant results for arithmetic scores suggesting food insecurity is more damaging to younger children than teenagers (Alaimo, Olson, and Frongillo, 2001).

In the case of studies using observation data, omitted variables correlated with diet and academic performance can bias the estimated effect of nutrition on achievement. As a result, more recent studies use more rigorous identification strategies to estimate the casual effect of nutrition on educational achievement. In 1995, the Minnesota state legislature

approved a grant providing free breakfast to a set of six treatment schools to test the effectiveness of extending a similar program statewide. In addition to the six treatment schools, three control schools were chosen to provide a natural set of counter-factual students. During the three years of treatment, each treated school reported an increase in math and reading achievement while the control schools' scores were relatively flat. In addition, teachers in treatment schools reported fewer students complained of headaches and stomachaches, students were more energetic, and had an easier time concentrating. Teachers also reported fewer disciplinary issues in the treated schools with morning disciplinary referrals declining between 15 and 50% for all treated schools (Wahlstrom and Begalle, 1999). The use of a set of control schools and the use of administrative data allow for a more plausible argument for causality.

A randomized control trial (RCT) in Jamaica selected seventh grade students in the lower-third of academic performance and randomized the students into a control group and two treatment groups. In one treatment group, students were given a school lunch and in the other treatment group, students were given a syrup drink. The study found that the students who received the school lunch performed better than the control and alternate treatment on an arithmetic test and had better attendance records. The results were also robust to controlling for attendance (Powell, Grantham-McGregor, and Elston, 1983). Another study in Jamaica randomized a breakfast treatment to rural students in grades 2 through 5. Treated students received a school breakfast while control students received one-quarter of an orange and an equal amount of attention. While treated students were shown to have improved in height and weight, significant arithmetic results were only apparent in the youngest of treated children (Powell et al., 1998). An RCT conducted in South Africa randomized 108 students into a treatment and control group and provided breakfast every school day to the treatment group for six weeks. The school breakfast was found to have a positive effect on cognitive performance for the treatment group (Richter, Rose, and Griesel, 1997).

In contrast to randomized control trials, the use of longitudinal data to study diet and academic achievement has become popular due to the ability to perform within-unit comparisons over time. While observational analysis is prone to omitted variable bias, panel data on children can eliminate time invariant unobserved heterogeneity that could otherwise bias observational studies. One particular study performed a randomized control trial and followed Guatemalan children during early childhood and through adulthood. Between 1969 and 1977, four villages in were randomly assigned a high protein drink and a low protein drink meant to be given to children between birth and 36 months of age. The children that were randomly chosen for treatment into the more nutritious drink

were tracked and interviewed in 2002. The study found positive effects of the intervention for the treated children. Treated women were found to have completed 1.2 more grades while both treated men and women had increases in both reading comprehension and non-verbal cognitive ability of one-quarter of a standard deviation (Maluccio et al., 2009).

In the case of the Guatemalan study, statistically and economically significant results can be found long after the end of treatment. This suggests that investments made to enhance nutrition in children at critical stages can have lasting effects on educational outcomes. On the other hand, an analysis of standardized test scores in Virginia show that interventions lasting as short as a week can have significant and immediate effects. This study identified schools that were under the threat of accountability sanctions if mean test scores were not improved. Researchers discovered that the school administrators systematically altered school lunch menus in an attempt to increase caloric counts during testing and finds that the schools who increased the caloric content of lunches the most saw the highest test score gains with an increase of 100 calories corresponding to increases of 7, 4, and 7 percentage points for mathematics, English, and social studies, respectively (Figlio and Winicki, 2005).

A more recent wave of studies estimate the effect of the National School Lunch Program and the rollout of School Breakfast programs across the U.S. Frisvold (2015) conducts a rigorous analysis of the School Breakfast Program using two identification strategies and multiple datasets and finds improvements in mathematics of 0.09 standard deviations and improvements in reading of 0.05 standard deviations for schools that adopt the School Breakfast Program. Similarly, Schwartz and Rothbart (2017) studies the impact of universal free lunch in New York City middle schools and the resulting impact on achievement. The study finds that an additional school lunch every two weeks improves math scores by around 0.08 standard deviations and improves language arts test scores by around 0.07 standard deviations. On the contrary, one recent study finds evidence of lower test scores when students are furthest away from the benefit receipt date. The study finds that when the students' family received SNAP benefits between 27 and 30 days prior to the test, math scores decline on average between 0.024 and 0.046 standard deviations (Cotti, Gordanier, and Ozturk, 2018). This suggests that families who receive federal food benefits may exhibit food insecurity at the end of the benefits cycle and that food insecurity may adversely impact academic performance.

While the literature covers a multitude of interventions spanning nutrition supplements, food stamps, school lunches, and school breakfasts, there are no studies (to my knowledge) covering interventions providing supplemental nutrition over the weekend for students reliant on free and reduced school lunches during the week. Since the in-

troduction of the School Lunch Program and Free or Reduced Lunch Program, families have become increasingly reliant on school lunches to provide adequate nutrition to children in low socioeconomic settings. In addition to school lunches, schools with a large proportion of free or reduced lunch eligible students tend to also provide breakfast to students. While these students receive a majority of their caloric intake from school during the school-week, the students must rely on their own household for nutrition over the weekend. The literature suggests that improving nutrition over the course of the weekend should positively impact student health, attendance, academic achievement, and behavior.

2.3 Background

During the 2011-2012 school year, a local non-profit conducted an intervention in two elementary schools in the Mississippi Delta. Third- and fourth-grade students at Brooks Elementary School in Bolivar County and third-, fourth-, and fifth-grade students at Stampley Elementary School in Coahoma County were selected for inclusion into the treatment. Students at these schools were overwhelmingly eligible for the Free and Reduced Lunch Program (FRLP) with 99% eligibility at Brooks Elementary and 94% eligibility at Stampley Elementary. According to questionnaires administered by the non-profit, food insecurity was prevalent at both schools with more than half of parents at the two schools reporting some degree of food insecurity. In addition to the intervention at Brooks Elementary and Stampley Elementary, the non-profit conducted another intervention at two more schools during the 2015-2016 academic year, however public test score data for the most recent intervention is not available at this time.

In order to be included in the intervention, students were required to return a permission form signed by a parent or guardian. In total, 174 students were included in the treatment (73 at Stampley and 101 at Brooks).[1] Each Friday between September 30, 2011 and May 18, 2012 recyclable bags filled with food were distributed to all students involved in the treatment. Each bag contained food intended to last for weekend consumption denoted in Table 2.1. In order for the student to receive the food, the student must have attended school on Friday. Food was not distributed over weeks in which students were on break including: Fall Break, Thanksgiving Break, Winter Break, and Spring Break.

[1]In 2011-2012, there were 56 third graders and 51 fourth graders who took the standardized tests at Brooks Elementary. At Stampley, there were 28 third graders, 34 fourth graders, and 38 fifth graders who took the tests.

Table 2.1: Contents of Weekend Supplemental Nutrition

2 bowls of cereal
2 small containers of fat-free milk
2 pieces fresh fruit
3 canned meats
1 cup applesauce
1 cup mixed fruit in syrup

2.4 Data

I use data from the Mississippi Department of Education that include performance metrics on the Mississippi Curriculum Test, 2nd Edition (MCT2). The MCT2 is the standardized test administered every spring to public school students in the state of Mississippi during the sample period. MCT2 tests students in Language Arts and Mathematics beginning in the third grade. The data for this analysis spans the academic years 2007-2008 through the 2011-2012 school years. MCT2 scores are reported for each public school in the state of Mississippi for each grade level. In addition to average scores, the data also include information about the distribution of test scores. This data include the percentage of students in each school-grade cell achieving between three thresholds creating four bins of test scores: Minimal, Basic, Proficient, and Advanced. These metrics allow for analysis on both the average and the distributional change in test scores as a result of the intervention.

I also obtained attendance records for all Mississippi public schools for kindergarten through fifth grade students for the sample years. These records contain the number of students absent from each school-grade cell on a daily basis throughout each school year. To create an outcome measure that is consistent across various school sizes, I compute the number of absences by day-of-week per enrolled student at the school-grade cell. Enrollment data comes from the Common Core of Data (CCD).

Lastly, I include control variables for changes in other programs that might also affect food insecurity. I merge data on Free and Reduced Lunch Program (FRLP) eligibility for each school-grade-year cell from the Common Core of Data and I include county level Supplemental Nutrition Assistance Program (SNAP) enrollment for each school. Table 2.2 reports summary statistics for outcome measures and demographics between the 2007-2008 and 2010-2011 school years.[2] Column 1 includes the full sample of Mississippi public schools that contain at least one of grades three, four, or five. Column 2 contains only

[2]Schools with missing information on the percent of students receiving Free and Reduced School lunches are omitted from the analysis.

the two treated schools, J.W. Stampley Elementary and Brooks Elementary. Comparing these columns reveals that the schools selected for the intervention score worse than the average Mississippi elementary school in both Language Arts and Mathematics test scores. The distributions for both tests follow a similar pattern with the treated schools having a larger percentage of students achieving Minimal and Basic standards and fewer students achieving Proficient and Advanced. Students at the schools selected for treatment also missed more days of school per student with those increases largely occurring on Mondays and Fridays. Lastly, the treated schools are populated by a larger share of Black students, have smaller average grade sizes, and have higher FRLP eligibility.

Table 2.2: Descriptive Statistics: Test Scores and Demographic Variables

Variable	(1) All MS Schools	(2) Treated Schools
LA Score	-0.048	-0.317
% Minimal	15.4	20.1
% Basic	36.2	45.5
% Proficient	37.7	27.9
% Advanced	10.7	6.4
Test Takers	87.6	42.7
Math Score	-0.046	-0.245
% Minimal	16.2	16.3
% Basic	28.5	37.9
% Proficient	45.1	41.6
% Advanced	10.2	4.3
Test Takers	87.6	42.7
Absences per student	6.00	6.35
Monday	1.31	1.37
Tuesday	1.17	1.21
Wednesday	1.11	1.19
Thursday	1.08	1.13
Friday	1.33	1.44
% White	41.4	0.0
% Black	54.5	98.4
% Hispanic	1.9	0.3
% Asian	0.5	0.0
% Free Lunch	77.5	98.3
Schools	532	2

The table above reports means for outcome variables (test scores) and demographic variables for the full sample of MS schools separately from the two treated schools. Outcome variables include normalized mean test scores in Langauage Arts and Mathematics on the MCT2 state test along with the number of test takers for each school-grade level. Demographic variables include the proportion of each school-grade that is white, black, Hispanic, or Asian as well as the percent of each school-grade cell that is eligible for Free and Reduced School Lunch. Demographic and FRLP data come from IPEDS while test score data come from Mississippi Department of Education.

2.5 Empirical Strategy

2.5.1 Identification

Previous qualitative surveys administered to teachers and parents of the treated students suggest that students were better behaved, more attentive in class, and more likely to attend school on Friday during the year of the intervention.[3] However, these impacts cannot be interpreted as casual effects of the intervention without considering how other students performed during the 2011-2012 school year. It is possible students that were not exposed to the intervention also performed better in school and had better attendance in the 2011-2012 school year for reasons unrelated to the intervention. Further, it could be the case that parents and teachers were more likely to answer positively on these surveys as a result of knowledge of the treatment in hopes of continued aid.

To help mitigate the concerns, I employ a difference-in-differences identification strategy to estimate the effect of the supplemental nutrition program on test scores and attendance for the students in the schools selected for the intervention. This design compares changes in outcomes of treated school-grade units to changes in outcomes for otherwise similar untreated school-grade units to control for any factors that might influence students in the same manner in the same years. If the necessary assumption, discussed below, is satisfied the estimated effect is the Intent to Treat (ITT) estimate of the intervention.[4]

The baseline specification for the test score outcomes follows a basic difference-in-differences framework:

$$y_{gst} = \gamma treat_{st} + \delta_g + \delta_s + \delta_t + \delta_s \times t + \varepsilon_{gst}, \tag{2.1}$$

where y_{gst} is a test score outcome measure for grade g at school s during academic year t. $treat_{st}$ is a binary variable which equals one for grades three through five at the two treated schools in the 2011-2012 academic year where γ is the estimated ITT. Also included are fixed effects for schools (δ_s), grades (δ_g), and academic years (δ_t). I also include a school specific linear time trend to capture differences in test score trends for each school. Standard errors are clustered at the school level to allow for correlation between grade levels within a school.

For the attendance outcomes, I can include one additional level of differences in the analysis. While standardized tests are only administered for students grade three and

[3] From summary report from the administering non-profit organization. Available upon request.

[4] Since not all students returned permission slips to enroll in the intervention, the effect estimated is not the Average Treatment Effect on the Treated.

above, the attendance data contains records for all students in kindergarten through fifth grade. I use this additional data to compare the absence rate *within* treatment schools across the treated grades (3-5) and the untreated grades (K-2). The DDD specification is similar to Equation (2.1):

$$y_{gst} = \gamma treat_{gst} + \delta_g + \delta_s + \delta_t + \delta_s \times t + \varepsilon_{gst}, \tag{2.2}$$

except the DDD coefficient, $treat_{gst}$, also varies by grade level since the sample also includes absence data for the untreated grades K-2.

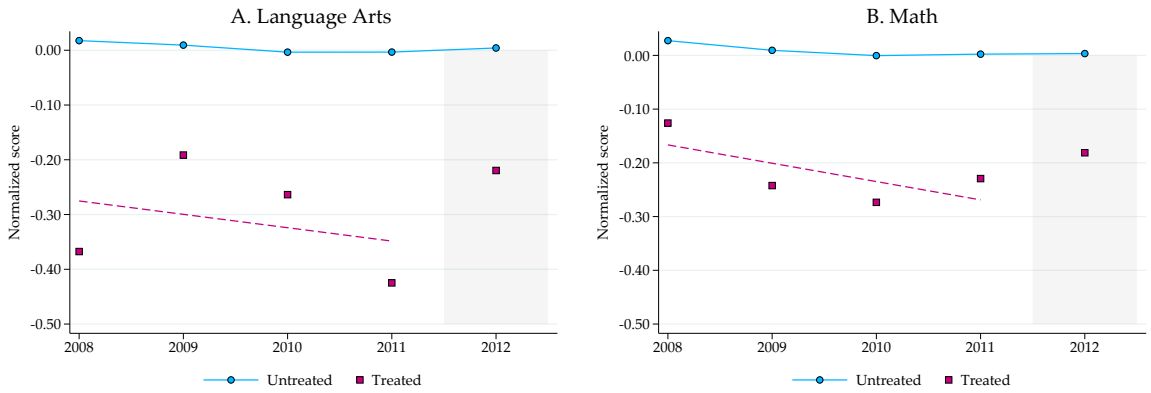
2.5.2 Selection of Comparison Group

As mentioned above, the estimated coefficient γ can only be interpreted as the ITT if the treated units would have evolved similarity to the control units in the absence of treatment. If the untreated units used as controls in the sample are systematically different from the treated units, the observed outcomes in the year of treatment cannot be reasonably used as a counter-factual outcome for the treated units. Thus, it must be the case that *changes* in test scores in the academic years prior to 2011-2012 are similar across the treated and untreated units. As described in Section 2.4, the schools selected for treatment had, on average, lower test scores, higher participation in FRLP, and had a larger share of Black students than the overall sample of schools in Mississippi. Figure 2.1 shows the trend in Language Arts and Mathematics test scores for the schools selected for treatment versus the sample of all other Mississippi Schools. Since the test scores for the full sample are normalized, the full sample of schools will be (near) mean zero for all years.[5] However, for the treated schools, the scores report where in the distribution of the sample the treated schools fall. Language Arts and Math scores for the treated schools are around 0.3 and 0.2 standard deviations lower than the mean, respectively. Additionally, both scores are slightly decreasing in the years leading up to the intervention. This trend could be the result of falling raw scores in the schools selected for treatment, gains in other schools, or both. Regardless, the trend in scores between the schools selected for treatment and the full sample of schools suggests that the full sample of schools might make for a suitable control group for the treated schools despite the level differences.

I also select a subsample of schools from the sample of all Mississippi public schools that is more similar to the two treated schools on the observables in which the treated schools most differ from the full sample: FRLP eligibility, size of grade level, and proportion of the student body who is Black. In addition, I also consider distance from the two treated

[5]Some schools are removed from the sample for missing data.

Figure 2.1: Mean Test Score Trends for Full Sample versus Treated Schools



Each figure above plots the mean test scores for the full sample of MS schools against the mean test scores for the treated schools in each year of the data. In addition, the red dashed line shows the trend in test scores for the treated schools in the three years before treatment. The shaded area, the 2011-2012 school year, denotes the year of treatment.

schools as-the-crow-flies. In many cases, matching (or inverse probability weighting) methods use a nonlinear model such as a probit or a logit model to find comparison units that are similar to the units selected for treatment. However, in this setting, a nonlinear model is not feasible since there are few schools selected for treatment and because the treated schools are largely boundary values for the selection criterion. Instead, I use Mahalanobis multivariate distance (MD) to rank the schools on similarity metrics which allocates more weight to schools that are more similar to the two treatment schools (Rubin, 1980). This algorithm takes selection variables and corresponding values on which to the match observations and calculates the distance between each observation in the sample from the mean values of the treated schools. The distance for each matching variable is weighted by the inverse of the variance-covariance matrix of the selection variables. The result is a weighted distance measure, measured in standard deviations, ranking how similar units are to the variable values on which the matching is constructed.

Table 2.3 lists the top 15 Mississippi schools with the highest Mahalanobis distance scores including the two treated school, Brooks Elementary and J.W. Stampley Elementary, in bold. The schools measured as most similar to the treated schools are all largely similar in the measured demographics by construction. In fact, there are many schools which are have smaller distance measures than the schools chosen for treatment since these schools better match the mean values of the variables than each individual school does alone. To create the matched sample of comparison schools, I select the subsample of comparison schools in top quartile in Mahalanobis distance so that the 25% of the sample most similar

to the treatment schools are chosen. Table 2.3 also details the bottom 15 schools selected in this subsample in terms of similarity to the treatment schools.

Table 2.4 repeats Table 2.2 adding the summary statistics for the schools in the comparison group determined by the constructed Mahalanobis multivariate distance measure. Selection on the three demographic characteristics and the average distance creates a comparison group that is much more similar to the treated units than the overall sample. The 133 schools in the comparison group only differ in FRLP eligibility by 4.2 percentage points and differ in the proportion of Black students by only 10.3 percentage points compared to the overall sample difference of 20.8 and 43.9 percentage points, respectively. The number of test takers for each grade differs by 14.6 students compared to the overall sample difference of 44.9. The more suitable control group generates average test scores that are also far more similar than the overall sample: average test scores in Language Arts are within 0.035 standard deviations and test scores in Math are within 0.021 standard deviations.

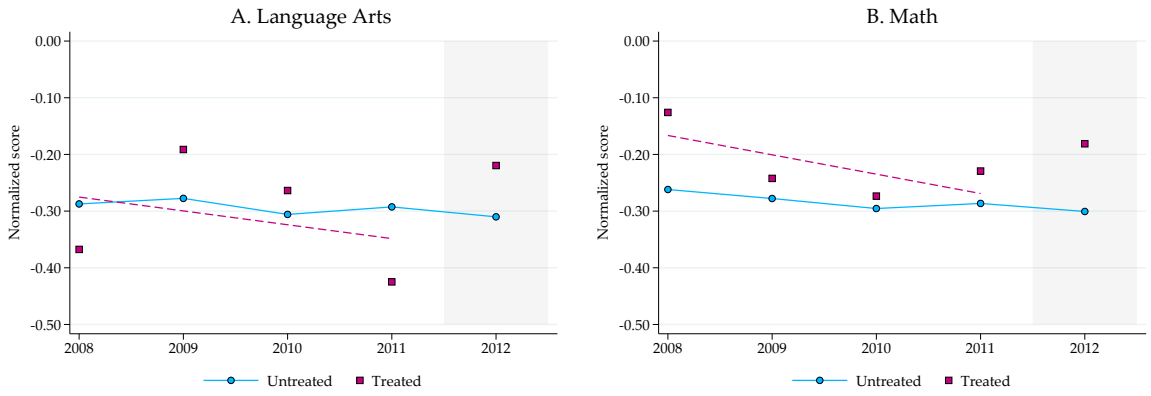
Table 2.3: Comparison Group by Mahalanobis Multivariate Distance

Rank	School	District	% FRLP	% Black	Test Takers	Avg. Dist	Mahalanobis Distance
1	Shelby School	North Bolivar	98.6	98.4	48.2	12.8	0.065
2	I T Montgomery Elementary	Mound Bayou	99.0	99.3	44.3	17.2	0.075
3	Myrtle Hall IV Elementary	Clarksdale	98.5	99.8	47.7	8.1	0.166
4	Lyon Elementary	Coahoma County	98.9	98.0	41.8	11.7	0.174
5	**J W Stamply Elementary**	Clarksdale	97.9	98.8	37.2	0.0	0.195
6	**Brooks Elementary**	North Bolivar	98.9	97.9	52.3	0.0	0.195
7	Booker T Washington	Clarksdale	96.8	99.0	45.3	7.7	0.224
8	Nailor Elementary	Cleveland	98.7	97.0	33.3	26.6	0.248
9	Hunter Middle	Drew	96.4	96.3	41.3	23.7	0.252
10	A W James Elementary	Drew	95.8	90.7	51.2	23.5	0.261
11	Geo H Oliver Elementary	Clarksdale	96.8	99.5	58.8	8.5	0.288
12	Jonestown Elementary	Coahoma County	98.9	99.5	45.5	18.3	0.299
13	McEvans School	Shaw	96.8	97.0	47.1	33.8	0.335
14	Friars Point Elementary	Coahoma County	97.9	98.0	30.9	17.5	0.335
15	West Bolivar Elementary	West Bolivar	95.9	92.8	65.4	27.9	0.346
121	Galloway Elementary	Jackson Public	96.1	99.0	41.8	126.6	1.844
122	Lake Elementary	Jackson Public	96.7	98.3	75.1	126.4	1.846
123	Philadelphia Elementary	Philadelphia	84.3	73.0	89.4	129.0	1.847
124	Sykes Elementary	Jackson Public	89.0	86.3	72.7	131.2	1.847
125	Hopkins Elementary	Jackson Public	96.0	98.3	81.9	124.8	1.854
126	Johnson Elementary	Jackson Public	94.7	98.8	56.6	125.6	1.858
127	Poindexter Elementary	Jackson Public	96.9	99.3	31.4	127.8	1.862
128	Lester Elementary	Jackson Public	94.4	96.5	48.2	129.6	1.874
129	George Elementary	Jackson Public	96.2	97.8	26.6	129.0	1.878
130	Noxapater Attendance Center	Louisville	80.6	50.3	30.0	120.6	1.880
131	Smith Elementary	Jackson Public	94.6	99.8	64.3	125.0	1.881
132	H W Byers Elementary	Marshall County	90.7	51.3	63.4	89.3	1.890
133	Key Elementary	Jackson Public	96.2	98.5	66.7	130.0	1.892
134	Green Elementary	Jackson Public	93.3	99.5	60.4	123.8	1.895
135	Houlka Attendance Center	Chickasaw County	78.7	39.5	43.1	94.9	1.907

The table above ranks the top 15 and bottom 15 schools selected in the subsample as determined by the Mahalanobis multivariate distance using the mean values of the two treated schools for %FRLP, % Black, number of Test Takers and average distance from Brooks and Stampley Elementary. Data on FRLP eligibility and % Black come from IPEDS while the number of test takers comes from the Mississippi Department of Education. Avg. Dist. is the average distance between the school and Brooks Elementary and Stampley Elementary and are computed using the Stata program geodist which calculates distance as-the-crow-flies. Treated schools in bold.

Figure 2.2 repeats Figure 2.1 with the comparison group selected using the Mahalanobis similarity score. This subsample better matches the *level* of test scores for both Language Arts and Math than the full sample but a worse job matching the *trend* in scores. In this figure, the potential treatment impact is more apparent with 2012 Language Arts and Math scores for the treated units outpacing the same year scores for the comparison group. This trend break suggests that the intervention had a positive impact on the treated schools in both subject areas. If the 2012 scores for the comparison group are used as a counter-factual for the treated units, this would suggest improvements in test scores larger than 0.1 standard deviations.

Figure 2.2: Mean Test Score Trends for Matched Schools versus Treated Schools



Each figure above plots the mean test scores for the subsample of MS schools most similar to the treated group as described in 2.5 against the mean test scores for the treated schools in each year of the data. In addition, the red dashed line shows the trend in test scores for the treated schools in the three years before treatment. The shaded area, the 2011-2012 school year, denotes the year of treatment.

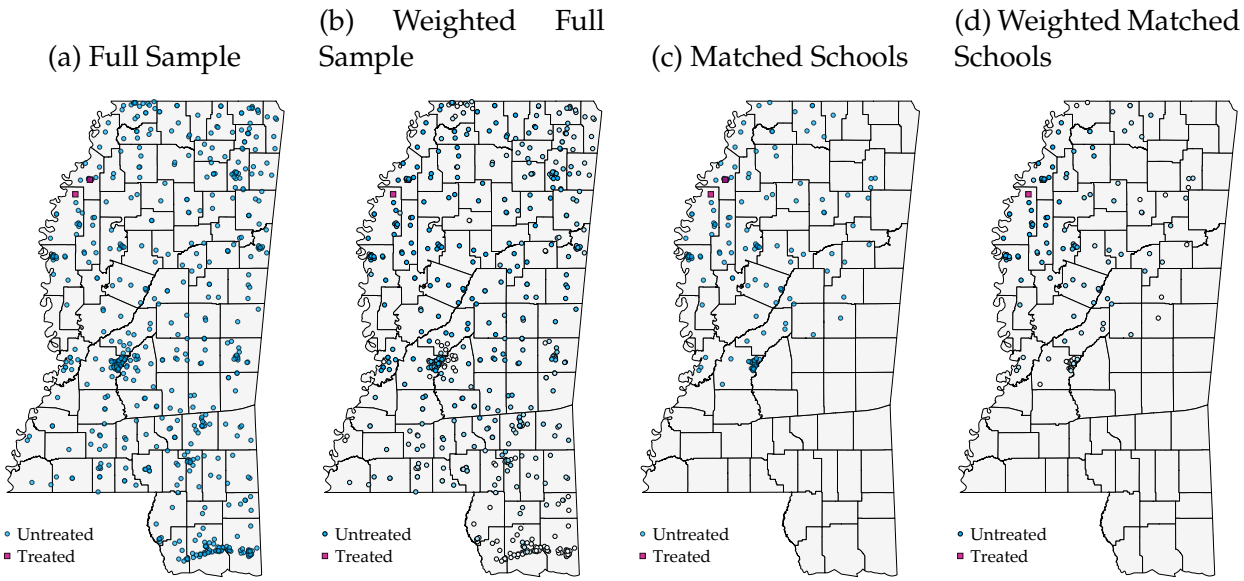
In the next section, I present results from the estimation of Equation (2.1) for the full sample and the selected comparison group in order to estimate the causal impact of the supplemental nutrition intervention on test scores. In addition, I also estimate Equation (2.1) for the full sample and the subsample using the Mahalanobis distance measure to weigh the observations. This specification allows for the schools most similar to the treated schools to take on more weight in the analysis than the schools that are less similar. Figure 2.3 shows the selection of the comparison groups in a map of Mississippi for each specification. In addition, the specifications that use similarly weights are shown in the varying shades of blue with the most similar schools in darker blue and the least similar schools in lighter blue.

Table 2.4: Descriptive Statistics: Test Scores and Demographic Variables

Variable	(1) All MS Schools	(2) MD Matched Schools	(3) Treated Schools
LA Score	-0.048	-0.282	-0.317
% Minimal	15.4	20.7	20.1
% Basic	36.2	41.7	45.5
% Proficient	37.7	31.2	27.9
% Advanced	10.7	6.5	6.4
Test Takers	87.6	57.3	42.7
Math Score	-0.046	-0.266	-0.245
% Minimal	16.2	21.5	16.3
% Basic	28.5	32.6	37.9
% Proficient	45.1	40.1	41.6
% Advanced	10.2	5.7	4.3
Test Takers	87.6	57.3	42.7
Absences per student	6.00	5.61	6.35
Monday	1.31	1.25	1.37
Tuesday	1.17	1.09	1.21
Wednesday	1.11	1.03	1.19
Thursday	1.08	1.01	1.13
Friday	1.33	1.24	1.44
% White	41.4	9.2	0.0
% Black	54.5	88.1	98.4
% Hispanic	1.9	1.2	0.3
% Asian	0.5	0.1	0.0
% Free Lunch	77.5	94.1	98.3
Schools	532	133	2

The table above reports means for outcome variables (test scores) and demographic variables for the full sample of MS schools, the upper quartile of MS schools most similar to the treated schools by Mahalanobis multivariate distance, and the two treated schools. Outcome variables include normalized mean test scores in Langauage Arts and Mathematics on the MCT2 state test along with the number of test takers for each school-grade level. Demographic variables include the proportion of each school-grade that is white, black, Hispanic, or Asian as well as the percent of each school-grade cell that is eligible for Free and Reduced School Lunch. Demographic and FRLP data come from IPEDS while test score data come from Mississippi Department of Education.

Figure 2.3: Geographical Depiction of the Four Variations of the Selection of Comparison Group



Each map above shows the selection of schools used for the comparison group in each corresponding specification in blue and the two treated school in pink. For the weighted specifications, the shaded blue fill is proportional to the school weights.

2.5.3 Inference

In the baseline specification, I cluster standard errors at the school level to allow for error correlations within the same school. However, recent research suggests that cluster robust standard errors perform remarkably poorly when there are few treated clusters (MacKinnon and Webb, forthcoming). As a result of the few number of treated units, the finite sample standard errors tend to be underestimated causing over-rejection of the null hypotheses. In this setting, there are only two treated schools and since schools are the unit of clustering, it is likely that the typical cluster robust standard errors are too small. To correct for this, I follow the RI-β algorithm prescribed by MacKinnon and Webb (forthcoming) by conducting a randomization inference exercise. I perform 3,500 placebo replications in which, instead of the observed treated schools, I suppose that two other random schools were chosen for treatment and I estimate Equation (2.1) under this supposition. This exercise compares the difference-in-difference coefficient estimated in Equation (2.1) with the distribution of placebo estimates. For each placebo replication, I draw two schools at random from the set of all schools in Mississippi with FRLP greater than 75%. This restriction of potential treatment schools is chosen so that only schools that might have been reasonably chosen for the intervention are included. I collect all

estimated placebo $\hat{\gamma}$ and generate two empirical p-values for each outcome variable. First, I use the two-sided test where I estimate the share of $\hat{\gamma}$ that are larger in absolute value than the $\hat{\gamma}$ estimated using the actual treated schools:

$$\bar{p} = \frac{1}{3500} \sum_{n=1}^{3500} 1 \cdot \left\{ |\hat{\gamma}_n| \geq |\hat{\gamma}_{true}| \right\} \tag{2.3}$$

In addition to the two-sided test, I construct a one-sided test that tests where there is an improvement in the outcome. For mean test scores and percent achieving Proficient and Advanced, the empirical p-value is

$$\bar{p} = \frac{1}{3500} \sum_{n=1}^{3500} 1 \cdot \left\{ \hat{\gamma}_n \geq \hat{\gamma}_{true} \right\}, \tag{2.4}$$

and for percent achieving Minimal and Basic and for absences, the empirical p-value is

$$\bar{p} = \frac{1}{3500} \sum_{n=1}^{3500} 1 \cdot \left\{ \hat{\gamma}_n \leq \hat{\gamma}_{true} \right\}. \tag{2.5}$$

Since MacKinnon and Webb (forthcoming) note that this algorithm has a tendency to under-reject, the one-sided test will provide a lower threshold for rejecting the null hypothesis of no improvement in outcomes. For each outcome, I present the clustered standard errors which are likely under-estimated along with the one- and two-sided tests using the empirical p-values.

2.6 Results

2.6.1 Standardized Test Scores

Table 2.5 reports the results from the estimation of Equation (2.1). Columns 1 through 4 report the results for Language Arts scores while columns 5 through 8 report the results for Math scores. The first two columns for each subject score include the full sample while the last two columns include only the subsample of the most similar quartile of schools based on the multivariate distance measure described above. Odd numbered columns report unweighted regressions while the even numbered columns report regressions that are weighted by the Mahalanobis distance measure. For each outcome, I report p-values using clustered standard errors in parenthesis along with one- and two-sided empirical p-values using RI-β from MacKinnon and Webb (forthcoming) in brackets. Columns 1 through 4 suggest improvements in Language Arts test scores regardless of specification.

The estimates using the full sample range from 0.24 to 0.28 from a pre-treatment mean of -0.317. Hence, the results suggest the intervention closed 75% to 88% of the gap in test scores between the treated school and the state average. The effect on math scores is slightly smaller at between 0.16 to 0.18 standard deviation improvement from a mean of -0.245 representing a closing of the test score gap of 65% to 73%. In total, the point estimates suggest significant test score improvements as a results of the nutrition intervention for the school-grade cells selected for treatment.

Table 2.5: Difference-in-Differences Estimates: Language Arts and Math Scores

	Language Arts Score				Math Score			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.240	0.241	0.284	0.276	0.166	0.161	0.177	0.174
CRVE p-value	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
RI-β 2-sided p	[0.184]	[0.203]	[0.157]	[0.158]	[0.378]	[0.388]	[0.235]	[0.237]
RI-β 1-sided p	[0.097]	[0.109]	[0.083]	[0.083]	[0.207]	[0.208]	[0.139]	[0.139]
Outcome Mean	-0.317	-0.317	-0.317	-0.317	-0.245	-0.245	-0.245	-0.245
Sample	Full	Full	Matched	Matched	Full	Full	Matched	Matched
Weights	No	Yes	No	Yes	No	Yes	No	Yes
Schools	523	523	132	132	523	523	132	132
Grades	3	3	3	3	3	3	3	3

Each column contains the estimated difference-in-differences coefficient where the outcome variable is either the standardized language arts or mathematics test score as denoted in the column title for a school-grade cell. Odd numbered columns are unweighted and even numbered columns use the normalized Mahalanobis distance measure as a regression weight. Columns One, Two, Five, and Six use the full sample of schools. Columns Three, Four, Seven, and Eight use the matched subsample of schools using the Mahalanobis distance measure. The Outcome Mean reports the mean value for the given outcome across the treated schools in the years prior to treatment. P-values using standard errors clustered at the school level are presented in parenthesis. One- and two-sided empirical p-values using RI-β are presented in brackets.

Across all eight specification, the p-values corresponding to each coefficient suggest the null hypothesis can be rejected at the 1% level. This is likely evidence that indeed the standard errors are under-estimated due to the low number of treated clusters in the sample. As such, these p-values should not be trusted and the RI-β empirical p-values are likely to be better estimates for inference. Using the 2-sided empirical p-values, we cannot reject the null hypothesis that the language arts treatment effect is not equal to zero. The p-values from the one-sided test lie just outside of the rejection region for a test of improving test scores. Additionally, the p-values for the mathematics scores are almost twice those of the languages arts scores and, as such, the null hypothesis also cannot be rejected. In total, the evidence suggests large improvements in test scores for the treated

units but the small number of treated clusters produces errors that make this conclusion difficult.

2.6.2 Distributional Treatment Effects

The discussion in the previous section presents evidence that average standardized test scores for the schools selected for treatment improved in the year of the intervention. The magnitude of these estimates are quite large and suggest that the treated students performed almost as well as the average Mississippi student despite having average test scores three-tenths of a standard deviation in previous years. However, due to the few number of treated units, inference on the average is difficult and the null hypothesis cannot be rejected by traditional methods.

On the other hand, the Mississippi Department of Education provides information on the distribution of students in each school-grade cell. Table 2.6 presents the difference-in-difference estimates for the percentage of students in each performance bin from worst (Minimal) in Panel A to best (Advanced) in Panel D for both language arts and mathematics. For both language arts and mathematics, there are declines in the percentage of students achieving at the lower threshold. The effect is larger again for language arts and the point estimate suggests a 42% improvement.[6] The reduction for mathematics is similar in proportional terms but smaller in magnitude. Again, the empirical p-values for the two-sided test are about twice as large for mathematics than for language arts. We are able to reject the null hypothesis of no improvement for language arts at the 10% level but the empirical p-value for mathematics is larger than 0.100 at 0.136.

The results for the second bin of achievement are mixed across language arts and mathematics. For language arts, the point estimates suggest declines around 11 percentage points while math achievement in this bin is around 3.5 percentage points larger. The point estimate for the percentage of students in the Basic bin is quite imprecise, however the reduction in students achieving Basic for language arts is significant at the 10% level for a two-sided test and at the 5% level for a one-sided test.

Since students move from one bin to another by construction, the decline in both Minimal and Basic for language arts suggest students are shifting into higher achievement. This result is reflected in the point estimates in Panel C for the percentage of students achieving Proficient standards. The point estimate is quite large at between 17 and 19 percentage points and the estimate is significant at the 5% level for the two sided test and the 1% level for the one-sided test using the empirical p-values. This suggests that

[6] 8.5 percentage point reduction off of a base of 20.1 percentage points

Table 2.6: Difference-in-Differences Estimates: Distributional Effects for Language Arts and Mathematics Test Scores

	Language Arts Score				Math Score			
A. Percent Minimal	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	-8.488	-8.441	-8.798	-8.720	-6.615	-6.482	-6.473	-6.440
CRVE p-value	(0.000)	(0.000)	(0.000)	(0.000)	(0.003)	(0.004)	(0.010)	(0.011)
RI-β 2-sided p	[0.112]	[0.138]	[0.108]	[0.108]	[0.256]	[0.287]	[0.193]	[0.195]
RI-β 1-sided p	[0.066]	[0.075]	[0.051]	[0.051]	[0.133]	[0.142]	[0.096]	[0.097]
Outcome Mean	20.147	20.147	20.147	20.147	16.284	16.284	16.284	16.284
B. Percent Basic								
Treatment	-11.661	-11.347	-11.237	-10.892	3.767	3.914	3.297	3.442
CRVE p-value	(0.000)	(0.000)	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)	(0.001)
RI-β 2-sided p	[0.075]	[0.085]	[0.072]	[0.078]	[0.475]	[0.488]	[0.273]	[0.268]
RI-β 1-sided p	[0.031]	[0.032]	[0.033]	[0.030]	[0.792]	[0.782]	[0.896]	[0.895]
Outcome Mean	45.511	45.511	45.511	45.511	37.858	37.858	37.858	37.858
C. Percent Proficient								
Treatment	19.222	18.848	17.804	17.624	-2.817	-3.059	-2.981	-3.087
CRVE p-value	(0.000)	(0.001)	(0.002)	(0.002)	(0.270)	(0.237)	(0.274)	(0.262)
RI-β 2-sided p	[0.013]	[0.016]	[0.029]	[0.029]	[0.641]	[0.660]	[0.337]	[0.337]
RI-β 1-sided p	[0.003]	[0.003]	[0.004]	[0.004]	[0.692]	[0.681]	[0.861]	[0.861]
Outcome Mean	27.884	27.884	27.884	27.884	41.595	41.595	41.595	41.595
D. Percent Advanced								
Treatment	0.925	0.935	2.220	1.977	5.699	5.662	6.189	6.117
CRVE p-value	(0.405)	(0.395)	(0.097)	(0.125)	(0.000)	(0.000)	(0.000)	(0.000)
RI-β 2-sided p	[0.803]	[0.808]	[0.347]	[0.353]	[0.193]	[0.196]	[0.107]	[0.107]
RI-β 1-sided p	[0.424]	[0.419]	[0.186]	[0.186]	[0.102]	[0.101]	[0.052]	[0.053]
Outcome Mean	6.447	6.447	6.447	6.447	4.268	4.268	4.268	4.268
Sample	Full	Full	Matched	Matched	Full	Full	Matched	Matched
Weights	No	Yes	No	Yes	No	Yes	No	Yes
Schools	523	523	132	132	523	523	132	132
Grades	3	3	3	3	3	3	3	3

Each column contains the estimated difference-in-differences coefficient. Each panel contains a different outcome variable where the outcome is the percentage of students in the school-grade cell achieving at a certain achievement threshold. The thresholds increase in achievement level from Minimal (lowest), Basic, Proficient, and Advanced (highest). Odd numbered columns are unweighted and even numbered columns use the normalized Mahalanobis distance measure as a regression weight. Columns One, Two, Five, and Six use the full sample of schools. Columns Three, Four, Seven, and Eight use the matched subsample of schools using the Mahalanobis distance measure. The Outcome Mean reports the mean value for the given outcome across the treated schools in the years prior to treatment. P-values using standard errors clustered at the school level are presented in parenthesis. One- and two-sided empirical p-values using RI-β are presented in brackets.

treated students are shifting from lower achievement bins to higher achievement bins for language arts scores. On the contrary, the point estimates suggest a reduction in students

achieving Proficient in math. This could be due to students leaving the Proficient bin and moving into either Basic or Advanced. However, this point estimate is very imprecise: even the under-estimated clustered standard errors are unable to reject the null hypothesis for these estimates.

Lastly, the effect of the intervention on the percentage of students achieving Advanced is presented in Panel D. For language arts, the estimates are economically small and are imprecisely estimated. However, the effect on mathematics is quite large relative to the baseline average for the treated schools in previous years. This large improvement in mathematics might explain the negative point estimates for the Proficient bin in Panel C: it may be the case that students previously achieving at Minimal are able to move to Basic and some students who previously achieved at Proficient are able to move into Advanced. Further, the point estimate is significant at the 10% level for the one-sided empirical p-value.

In total, analysis on the distribution of test scores suggest reductions in students achieving at lower thresholds with shifts towards higher achievement bins. For language arts, these improvements largely stem from students exiting the Minimal and Basic bins and entering the Proficient bins. For mathematics, there is some evidence students leave the Minimal bin and other students enter the Advanced bin but without micro-level data, it is not possible to track which students move across which bins.

2.6.3 Attendance

In addition to test score outcomes, I also explore how daily attendance is affected by the supplemental nutrition intervention. As mentioned above, attendance data is available for students in kindergarten through fifth grade. As a result, this data allow for a third dimension of comparison in which untreated students (K-2) in treated schools can be compared to treated students (3-5) in treated schools. The results from this section can help to explain the mechanisms by which the nutrition intervention improves test scores. If the supplemental nutrition program indeed improves nutrition over the weekend, students should feel healthier on Mondays and should be less likely to miss school on Mondays. On the other hand, the intervention could also act as an in-kind transfer in which students must attend school on Fridays to receive the transfer. If this is the case, attendance on Fridays should improve.

Table 2.7 reports the results of this strategy where each panel separately reports the estimated impact of the intervention on the number of absences per enrolled student for each day of the week in the year of the intervention. Panel A reports these estimates for

Mondays. On average, students in the treated schools missed around 1.53 Mondays per year making Monday the most frequently missed school day of the week. The estimates suggest a reduction in Mondays missed in the year of treatment by 0.18 to 0.21 days per enrolled student representing a 11-14% decline in Monday absenteeism. These estimates are also precisely estimated with the one-side p-value significant at the 10% level. The results in Panel B are similar to those in Panel A with declines in attendance around 15-16% These estimates are somewhat less precise with the one-side empirical p-value near the cutoff of the rejection region at 0.109.

On the other hand, the results in Panels C and D suggest no change in attendance for Wednesday or Thursdays. Both estimates are economically small and not precisely estimated across all specifications.

Lastly, the effect on Friday attendance seems to be the largest across the week. The reduction in absenteeism on Fridays is between 20-23% and is significant at the 5% level for both one- and two-sided tests. These results lend evidence to the mechanisms presented above. The largest effect of the intervention on attendance is due to the transfer effect on Fridays. However, Monday is the most frequently missed day of school for the treated units and the nutritional intervention reduces absenteeism on Mondays and Tuesdays by around 15%.

Table 2.7: Triple Difference Estimates: Absences per Enrolled Student by Day-of-Week

A. Monday	(1)	(2)	(3)	(4)
Treatment	-0.214	-0.208	-0.177	-0.175
CRVE p-value	(0.033)	(0.037)	(0.073)	(0.076)
RI-β 2-sided p	[0.134]	[0.158]	[0.111]	[0.110]
RI-β 1-sided p	[0.052]	[0.063]	[0.040]	[0.040]
Outcome Mean	1.533	1.533	1.533	1.533
B. Tuesday				
Treatment	-0.202	-0.206	-0.207	-0.202
CRVE p-value	(0.004)	(0.003)	(0.004)	(0.005)
RI-β 2-sided p	[0.253]	[0.273]	[0.162]	[0.172]
RI-β 1-sided p	[0.118]	[0.134]	[0.071]	[0.068]
Outcome Mean	1.318	1.318	1.318	1.318
C. Wednesday				
Treatment	-0.042	-0.044	-0.046	-0.042
CRVE p-value	(0.008)	(0.006)	(0.037)	(0.057)
RI-β 2-sided p	[0.839]	[0.827]	[0.358]	[0.347]
RI-β 1-sided p	[0.516]	[0.503]	[0.751]	[0.749]
Outcome Mean	1.333	1.333	1.333	1.333
D. Thursday				
Treatment	0.000	-0.003	0.000	-0.001
CRVE p-value	(0.999)	(0.964)	(0.994)	(0.991)
RI-β 2-sided p	[0.808]	[0.814]	[0.333]	[0.338]
RI-β 1-sided p	[0.353]	[0.352]	[0.122]	[0.123]
Outcome Mean	1.187	1.187	1.187	1.187
E. Friday				
Treatment	-0.351	-0.342	-0.311	-0.304
CRVE p-value	(0.000)	(0.000)	(0.000)	(0.000)
RI-β 2-sided p	[0.039]	[0.051]	[0.036]	[0.035]
RI-β 1-sided p	[0.017]	[0.023]	[0.028]	[0.028]
Outcome Mean	1.504	1.504	1.504	1.504
Sample	Full	Full	Matched	Matched
Weights	No	Yes	No	Yes
Schools	523	523	132	132
Grades	6	6	6	6

Each column contains the estimated difference-in-differences coefficient where the outcome variable is the number of absences on each day of the week per enrolled student. The third difference in this specification uses the untreated kindergarten through second grade students. Odd numbered columns are unweighted and even numbered columns use the normalized Mahalanobis distance measure as a regression weight. Columns One and Two use the full sample of schools. Columns Three and Four use the matched subsample of schools using the Mahalanobis distance measure. The Outcome Mean reports the mean value for the given outcome across the treated schools in the years prior to treatment. P-values using standard errors clustered at the school level are presented in parenthesis. One- and two-sided empirical p-values using RI-β are presented in brackets.

2.7 Conclusion

In this paper, I estimate how improved nutrition affects elementary school students living in the most food insecure region in the United States. Although the total expenditure per student is relatively small, the estimated improvements in test scores and attendance are quite large. I find that students attending the school selected for treatment had larger, albeit imprecise, average test scores in both language arts and mathematics. The improvements in average scores is higher for language arts than for mathematics, but the intervention was able to close the gap in test score by around 70% from the state average.

I also decompose the effect of the intervention on the distribution of test scores by using the percentage of students achieving at various levels as outcomes. For language arts, I find large reductions in the percentage of students achieving at the lowest two bins and large increases in the percentage of students achieving at the Proficient bin. I find no effect of the intervention on moving students in the Advanced bin for language arts.

The results for mathematics are smaller but perhaps more divergent. While I find small reductions in those achieving at the lowest bin, I am unable to trace movements into the middle two bins. The estimates are smaller and less precise and do not represent traceable shifts from the lower tail. On the other hand, I do find large improvements relative to the baseline in the percentage of students achieving Advanced in mathematics that might be driven by a reduction in the students achieving at Proficient. These improvements in the highest achievement threshold are the most precisely estimated effects for mathematics scores.

Lastly, I also find improvements in attendance for students at the schools selected for treatment, but only for those in the treated grades. I use daily administrative attendance records to track changes in attendance by day of the week. The results provide evidence for the mechanisms by which test scores are improved. First, I find the largest effect of the intervention on attendance on Fridays with reductions in absenteeism by 20%. This evidence is consistent with the transfer program mechanism by which students must be present at school on Friday to receive the nutrition bundle. However, I also find significant improvements in attendance on both Monday and Tuesday. For the set of selected schools, Monday was the most commonly missed day of school prior to the intervention and attendance increased by over 10% on Mondays and almost 15% on Tuesdays. This is evidence that indeed the students selected for treatment had better nutritional intake over the course of the weekend subsequently improving general physical condition at the beginning of the week.

The results from this paper add to the literature of the role nutrition plays in human

capital accumulation. The effect sizes from this one-shot intervention are on the higher end of the distribution from the education treatment effects literature. Further, these gains were achieved at a rather low cost. Kraft (2018) categorizes an intervention as "Large Effect Size/Low Cost" if the effect size is greater than 0.2 standard deviations and the cost is less that $500 per pupil. Using the estimated improvement in language arts test scores, this intervention meets the benchmark for a large effect size. Further, the cost of the intervention must be below $16.67 per week per student to be categorized as low cost. At an estimated $3.63 per food bundle[7] plus administrative costs, it is very likely this intervention remains beneath the $500 per student cost over the course of the 30-week school year. Lastly, this intervention is likely quite easy to scale. Since schools already have sufficient infrastructure to provide breakfast and lunch during school days, these items can be bought in bulk with the existing meal orders and distributed to students through the existing school infrastructure. As a result, this intervention is very likely to achieve the definition on Easy to Scale from Kraft (2018). However, the setting of this intervention is important when considering the external validity of these treatment effects. Since the Mississippi Delta region is the most food insecure region in the United States, it might be unreasonable to expect similar effect sizes if the intervention were to be repeated in another setting. Students that face a lower degree of food insecurity are unlikely to respond in a similar manner as the students in this setting. However, increasing the scale of this intervention across the Mississippi Delta is likely to achieve similar results for elementary school students in the area.

[7]This estimate is arrived at by assuming the following unit costs: cereal $0.33, milk $0.25, fruit $0.25, canned meat $0.50, applesauce $0.23, and mixed fruit $0.50.

Chapter 3: The Marginal Congestion of a Taxi in New York City

3.1 Introduction

The congestion of transportation infrastructure has both macro and microeconomic costs. Transportation affects the organization and scale of economic activity in cities (Redding and Turner (2015); Ahlfeldt et al. (2015)): while agglomeration saves transportation costs, these savings diminish as density congests infrastructure. When transportation infrastructure becomes congested, residents both bear and impose external costs. In the context of road congestion, costs include the value of time (Anderson (2014)), pollution and health (Currie and Walker (2011); Knittel, Miller, and Sanders (2016)), carbon emissions (Barth and Boriboonsomsin (2008)) and subjective well-being (Kahneman and Krueger (2006); Anderson et al. (2016)).

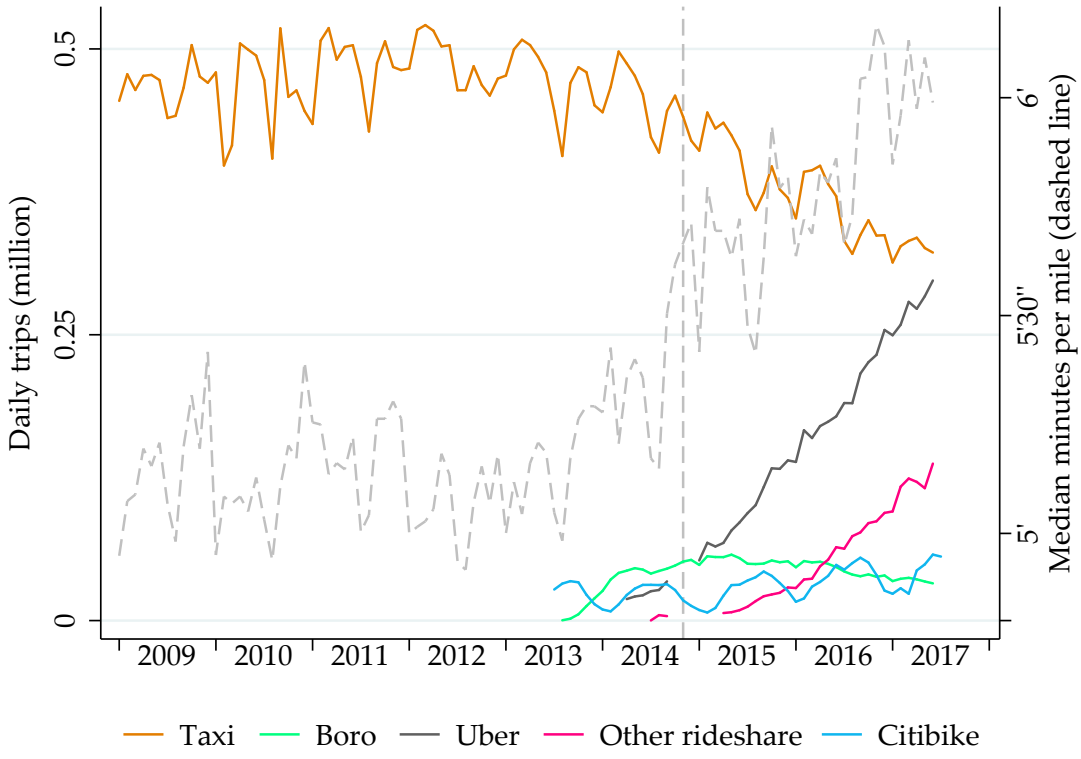
We document that traffic in New York City, one of the most productive cities in the United States (Hsieh and Moretti (2017)), has slowed down substantially since 2013 (see Figure 3.1). Taxi trips within midtown Manhattan have slowed down on the order of 15%, but the slowdown is widespread: it shows up in taxi trip records throughout the day and over the entire city, on highways – as measured though a novel dataset that we assemble from EZ-pass sensors – and in responses to the American Community Survey.

Cities such as New York are undergoing technological and policy changes with uncertain costs and benefits: roads may be under increasing demand from transportation network companies (also known as ridehail) and parcel delivery services, while space in the city is increasingly allocated to bikeshare, bike lanes and pedestrians (Sadik-Khan and Solomonow (2016)).

We evaluate a policy change that is unique to New York City: starting in August 2013, the city partially deregulated its medallion taxi industry by authorizing a new class of taxi. "Boro" taxis are almost identical to the city's 13,237 traditional yellow medallion taxis, but are painted green and are restricted from picking up passengers at airports or in Manhattan south of East 96th street and West 110th Street. We exploit the unique natural experiment provided by the roll-out of up to 6,539 new taxis searching for passengers under a well-defined spatial constraint to causally estimate the impact of taxi supply on traffic congestion in New York City.

We employ a dataset with 1.3 billion taxi trip records from 2009 to June 2016, which contain precise times and geographic coordinates of pickups and dropoffs, to construct a dataset of street-level speed that spans both sides of the exclusion boundary in northern Manhattan and the period prior to and after the roll-out of boro taxis. Our measures of

Figure 3.1: Daily Trips in New York City by Selected Transportation Modes, and Yellow Taxi Median Minutes per Mile. Monthly Statistics 2009-2016



Note: Trips from ridehail providers such as Uber and Lyft are available only for Q2 and Q3 in 2014, and from 2015 onwards. Dashed gray line plotted on second axis is the median minutes per mile among all yellow taxi trips in a given month. Vertical dashed line represents the Nov 7, 2014 date on which speed limits were lowered to 25 mph on most city streets under the "Vision Zero" traffic safety initiative.

speed are constructed at a high spatial resolution by selecting the trips that involved a pickup and a dropoff along the same avenue (about 69,300 trips per month in our area of interest) and either averaging or projecting travel times onto street segments or flexible functions.

Avenues flow faster toward the north of Manhattan, as density and traffic decline, and all avenues have slowed down over time. Following the introduction of boro taxis, however, a stark speed gap opened up north of the boro taxi exclusion boundary (see Figures 3.2 and 3.3). In our baseline estimates we find that the boro program's roll-out caused an overall 8-9% slowdown in the speed of traffic in northern Manhattan.

Our estimates are robust to controlling for observable changes in road use (such as the deployment of bike lanes and bikeshare, among others), as well as a large set of specifications to address unobservable confounders. Unobservables that could be a cause

for concern include differential trends over the city's geography (for instance, due to economic recovery or gentrification) or a differential impact from the arrival and growth of ridehail as a transportation alternative. Estimates of the boro program's impact track the timing of the roll-out of taxi supply, which ramped up over the last quarter of 2013, and are identified under narrow time windows around the roll-out. Our estimates are also robust to trends by avenue and treatment status that are linear in time but are also allowed to kink and track the rapid growth in ridehail. We also find that our results are not sensitive to differential time trends over space in the measurement error of geographic coordinates, a potential confounder in research that employs data from GPS units within dense urban settings. We address spatial spillovers, which could arise naturally as boro taxis transport passengers south of the pickup exclusion boundary, and congestion could be transmitted across the boundary through gridlock: alternative estimation strategies (which include a triple-difference interaction that exploits the direction of traffic, as well as spatial econometric models allowing for links between street segments) are consistent with the upper range (9%) of our impact estimates.

We next quantify the impact of boro taxis not in terms of the deregulation episode as a whole, but in terms of taxi supply. As a benchmark, we employ the taxi trip data to measure taxi activity in pickups and instrument for pickups with the boro taxi roll-out; we then rescale a congestion elasticity estimated in terms of pickup units to supply using a back-of-the-envelope proportionality assumption. Recent research on the taxi industry, however, has emphasized that pickups are the observed equilibrium outcome of search between unobserved demand and supply (Buchholz (2017); Frechette, Lizzeri, and Salz (2016)).[1] In our context, this implies that taxi pickups and supply may not be in proportion to each other.

We address the observability of taxi supply by collecting a sample of 29 aerial orthoimagery scenes going back to 2010. We digitize the location of 132 thousand yellow taxis, boro taxis and other vehicles in the imagery. Our data allows us to measure changes to the patterns of vehicle density over space, comparing the period before and after the roll-out of boro taxis.[2] Our data can be used to estimate the relationship between supply

[1]See also Lagos (2003) for an earlier analysis using a survey of taxi trip sheets (an earlier generation of taxi trip record data, hand-recorded by drivers with coarse location identifiers). The taxi record data we employ has been used to study labor supply (Farber (2015)), moral hazard in contracting (Schneider (2010); Jackson and Schneider (2011)), behavioral cues (Haggag and Paci (2014)) and learning (Haggag, McManus, and Paci (2017)), but not to estimate congestion externalities in transportation.

[2]To our knowledge, ours is the first application of vehicle counts from aerial imagery in economics, or to the study of the long-run impact of regulation on vehicle density and congestion in any discipline. See Donaldson and Storeygard (2016) for a discussion of satellite imagery in economics. Traffic engineers have employed aerial imagery for decades (e.g. Johnson (1928); Greenshields (1948)) for infrastructure design studies and more recently to calibrate agent-based models of traffic flow. A recent literature in

and taxi matches. We use it to directly measure changes in taxi supply at a high degree of spatial resolution, which we relate to a set of heterogeneous speed impacts estimated over the boro taxi zone. Heterogeneity in impacts is related to baseline traffic density. We obtain a distribution for the ratios of changes in travel time to changes in supply, i.e. congestion elasticities. We plot our measured elasticities against baseline taxi density to non-parametrically estimate a marginal congestion curve that is identified from the supply shock produced in northern Manhattan by the boro taxi program. We find that congestion elasticity from taxi supply rises and is slightly convex in density, reaching 0.30 at the higher range that best extrapolates to denser areas in midtown Manhattan.

Lastly, we apply our estimated congestion elasticity curve with a study of New York City's slowdown since 2013. We document the slowdown using taxi meter data and by contributing two new sets of time series to the state of knowledge on traffic speed and volumes in New York City: i) archived highway speeds reported from a network of EZ-pass sensors, and ii) vehicle miles traveled by vehicle class, estimated from odometer inspections. New York City policy makers have failed to find a link between growth in the supply of ridehail and the city's slowdown. As we discuss below, the city's congestion studies do not employ the available data exhaustively and fail to meet the standards of modern empirical research.

We digitize aerial imagery samples in an area of midtown Manhattan and find a 14.1% increase in vehicle density since the summer of 2013 over the prior three years. We also detect a substantial change in the observed color composition of (non-taxi, non-truck) cars in midtown Manhattan: whereas 29.3% of cars counted in overhead imagery prior to the summer of 2013 were black, that share has since increased to 45.7%. VIN-level records from the New York state Department of Motor Vehicles indicate that while 20.9% of privately owned cars in New York City are black, 73.0% of For-hire vehicles (FHVs, a category that includes ridehail) are black. Employing color-share data, we estimate that the density of for-hire vehicles in midtown Manhattan has increased by 222.9% since the summer of 2013. We also estimate a 6.8% decline in the number of private cars on the road.[3] Taking a 15.2% median slowdown in midtown Manhattan and a 0.30 estimated congestion elasticity as reference values, we find that the 30.5% increase in the joint supply

machine learning addresses the detection of vehicles from aerial images (e.g., Reilly, Idrees, and Shah (2010), Mundhenk et al. (2016); Razakarivony and Jurie (2016)) but its output has yet to be applied in empirical research.

[3]Aerial images combined with color-share data allow us to separately estimate the density of vehicles other than taxis, such as private cars and ridehail. An advantage of our approach is that we can measure vehicle density in dense, downtown city streets. Previous studies using counts from loop detectors, stationary traffic cameras or gantries (e.g. Xie and Olszewski (2011); Li, Purevjav, and Yang (2017)) are usually limited to highways or arterial roads.

of taxis and FHVs in midtown Manhattan over the past three years (taxis down 11.8%, FHVs up 222.9%) can account for 61.8% of the slowdown in traffic. Overall, evidence from aerial counts and the spatial and time series patterns of New York City's slowdown do not appear to be consistent with rival explanations such as bike lanes, construction or use by cyclists, pedestrians or trucks.

Section 3.2 below describes the boro taxi program. In Section 3.3 we use taxi trip records to build historical street speed data. In Section 3.4 we estimate the impact of the boro program on traffic speed. In Section 3.5 we derive a congestion elasticity curve for New York City in terms of taxi supply. In Section 3.6 we apply our estimated congestion curve to the issue of increased vehicle density due to supply growth from ridehail. Section 3.7 concludes.

3.2 Background

Taxi service in New York City is regulated by the Taxi and Limousine Commission (TLC). Among other regulations, the TLC sets fares, licenses drivers and cars and issues medallions, which are a fixed number of permits that must be attached to a licensed car for it to operate as a taxi (i.e. accept street hails) in the five boroughs of New York City. The number of medallions was fixed at 11,787 for decades, with a few small increases after 1996, and remained at 13,237 between May 2008 and August 2013. The only type of medallion taxi available up to August 2013 was the traditional yellow taxi. Although yellow taxis are statutorily allowed to pick up passengers in all five boroughs of New York City, most taxi pickups occurred in Manhattan, particularly midtown, and at the city's two airports. The spatial distribution of yellow taxi pickups is illustrated in Figure 3.12 using data for June 2013. Areas outside of northern Manhattan and the remaining four boroughs had limited taxi coverage, and throughout most of New York City residents could not dependably hail a yellow taxi from the curb. Prior to ridehail apps the most common method for residents in the outer boroughs to hire car transportation was to call radio-dispatch bases for pre-arranged transportation from "for-hire vehicles," as well as unlicensed street-hails offered by some for-hire vehicles and other informal providers.[4]

In August 2013 a new class of taxi was allowed on the road: painted green and referred to in regulation as "street-hail livery" (SHL) taxis, or alternatively as "boro" or green taxis, these taxis and their drivers are regulated like yellow taxis with a major distinction: passenger pickups (whether pre-arranged or street-hails) are not allowed at the two city airports or south of East 96th Street and West 110th Street in Manhattan, an area referred to

[4]See Gao Hodges (2012) for the history of the city's taxi cab industry, including unlicensed activity.

as the "hail-exclusion zone." The objective of the boro taxi program was to increase street-hail availability in the outer boroughs and provide a regulated alternative to unlicensed street-hails.[5] Following a multi-year political and regulatory process, the roll-out of the boro program involved an initial sale of 6,000 medallion permits, as well as a fleet roll-out and driver recruitment process carried out by medallion owners. Additional permits were put on sale in August 2014. Figure 3.4 contains a timeline of the number of unique boro taxis that recorded at least one trip during each month.

3.3 Data

There is no publicly available dataset that contains information on street speed at a high level of spatial resolution in New York City and over a historical period going back prior to the roll-out of the boro taxi program.[6] In this section we describe how we employ historical records of taxi trips from the Taxi and Limousine Commission (collected from GPS-enabled taxi meters since 2009 under the "Taxicab Passenger Enhancement Program") to construct measures of historical street speeds, using three alternative methodologies. In the next section we employ these measures as outcome variables to quantify the impact of the roll-out of the boro program over space and time.

3.3.1 Average Travel Times at a 10 Meter Resolution

The simplest approach to measuring the speed of traffic, and the approach we favor throughout the paper, is to average the rate of travel across all taxi trips for which we are confident that the taxi traversed a certain interval over a certain period. To estimate average travel times at a specific time and place, for each month in the data we select the taxi trips that were entirely contained within a rectangular strip running along one of the north-to-south avenues of Manhattan, and in which the pickup and drop-off locations are oriented in the direction of traffic. On avenues with two directions of traffic, such as Park Avenue, we classify trips according to direction. We refer to each avenue and direction of traffic as a "run."

[5] See www1.nyc.gov/office-of-the-mayor/news/362-13/mayor-bloomberg-more-1-000-boro-taxis-now-new-york-city-street for a NYC government press release on the roll-out of the program and its intended goals. Other differences with yellow taxis are that 20% of boro medallions require the taxi to be wheelchair-accessible, and the taxis are required to undergo vehicle inspections at half the regularity of yellow taxis.

[6] Recent studies of traffic congestion have employed probe data from cell-phones, obtained from the Google maps API (e.g. Hanna, Kreindler, and Olken (2017); Akbar and Duranton (2017)), but historical data was not available to us over our period of study. Additionally, travel time APIs are developed with the aim of prediction and their methodologies may changed in ways that render them not comparable over long periods or over space. Another alternative, the INRIX dataset, is constructed from probes on fleets of commercial vehicles and is not as attractive as our methodology for dense urban settings such as New York City.

Every trip record contains the time as well as the latitude and longitude for each pickup and drop-off, as provided by a GPS-enabled taxi meter. We project the coordinate of each endpoint perpendicularly onto a straight line down the center of each avenue, allowing us to locate each trip endpoint on this line. We then calculate for each trip the ratio of the travel time to the line distance, to obtain each trip's average seconds per meter (i.e. the inverse of speed). To calculate average seconds per meter at a 10 meter resolution, we subdivide each run into 10 meter intervals (or "bins"), and for each of these average the travel time per meter of all the taxi trips that fully traversed the interval. Variation in the average travel time between bin i and bin $i + 1$ therefore comes from the variation in the average travel times of taxi trips that fully traverse bin i but not bin $i + 1$, or vice-versa.[7]

Since our objects of interest are (averages of) speed over a particular segments, the averages obtained by the method described above could be subject to biases due to aggregation as well as changes in trip selection over space and time. As an example of this potential concern, consider two 10 meter bins in a relatively uncongested area in the north of the city, and assume traffic moves at the same speed along both bins. Suppose the northernmost of these bins contains a hotel driveway, whereas its neighbor to the south does not. The bin containing the hotel originates many long trips to a congested downtown area, whereas the neighboring bin originates short trips within a relatively uncongested interval. In this example, the average taken over the travel time of the trips crossing each bin will be upwardly biased as an estimator of the true marginal travel time of crossing each bin due to congestion that affects the taxis traveling downtown from the hotel (an aggregation bias) and, further, the bias will be larger for the bin containing the hotel since its average does not contain the short, local trips originating in the southern neighbor (a selection bias). Aggregation bias can also interact with time, as the bin may be located next to a nightclub or park, and trips may originate at times when traffic is lower, such as nights or weekends.

A third type of potential bias could arise from the type and behavior of the "probe" that provides a taxi trip record, e.g. whether a car is a yellow or boro taxi. We employ data on both types of taxis, since boro taxis increase the sample in the north of the city, but for this reason and because boro taxi trips are by definition only possible after the launch of the boro program, variation across taxi type in driving speed would bias our travel time estimates in a manner that could violate parallel-trend assumptions required for difference-in-differences evaluation. To investigate this possibility, we exploit the

[7]The within-avenue sample is presented in Table 3.1. The outcome of the data construction step is plotted in Figures 3.2 and 3.3. For ease of presentation, it is averaged again at the annual level, and expressed in miles per hour, rather than seconds per meter. Observable changes in speed are discussed in Section 3.4 below.

Table 3.1: Seconds-Per-Meter for Within-Avenue Taxi Trips, Northern Manhattan, Jan 2009 to Jun 2016

	N	Mean	Median	Std Dev	Min	Max
West End Northbound	262,340	0.223	0.199	0.127	0.024	6.37
West End Southbound	379,378	0.200	0.181	0.112	0.027	8.21
Broadway Northbound	1,068,259	0.205	0.181	0.130	0.026	10.90
Broadway Southbound	1,027,907	0.212	0.187	0.134	0.025	11.18
Amsterdam Northbound	514,332	0.148	0.125	0.119	0.029	12.55
Amsterdam Southbound	93,688	0.238	0.204	0.166	0.040	13.40
Columbus Northbound	1,308	0.338	0.230	0.370	0.080	4.48
Columbus Southbound	317,496	0.186	0.146	0.174	0.033	10.97
Manhattan Northbound	4,680	0.315	0.231	0.339	0.076	5.48
Manhattan Southbound	9,878	0.245	0.196	0.231	0.073	5.88
CPW Northbound	132,153	0.224	0.181	0.204	0.052	8.24
CPW Southbound	132,674	0.230	0.189	0.203	0.039	8.85
5th (S)	199,727	0.242	0.191	0.209	0.035	11.26
Madison (N)	347,121	0.209	0.163	0.190	0.035	7.74
Park Northbound	188,904	0.238	0.196	0.193	0.026	10.16
Park Southbound	157,976	0.281	0.224	0.243	0.030	8.39
Lexington (S)	272,852	0.219	0.172	0.201	0.033	13.22
3rd (N)	440,152	0.155	0.121	0.153	0.031	14.69
2nd (S)	427,408	0.194	0.153	0.174	0.026	10.07
1st (N)	260,863	0.157	0.119	0.160	0.031	10.29
Upper East Side	2,295,003	0.201	0.157	0.189	0.026	14.69
Upper West Side	3,944,093	0.201	0.177	0.142	0.024	13.40
All	6,239,096	0.201	0.171	0.161	0.024	14.69

Excluded: Riverside Drive, FDR Drive, Adam Clayton Powell Jr. Boulevard, Malcolm X Boulevard. Statistics are for the seconds per meter, calculated as travel time over the distance along an avenue centerline. The following are statistics in miles per hour for all within-trip observations: the average of 0.201 seconds per meter is the inverse of a speed of 11.13 miles per hour. The maximum is 93.2 miles per hour, the minimum is 0.15 miles per hour, and two standard deviations above the mean is 4.27 miles per hour. Not described in the table above, the 1st and 99th percentile seconds per meter are 0.0789 and 0.762, which are the inverses of 28.4 and 2.93 miles per hour, respectively.

Table 3.2: Descriptive statistics

	Mean	Median	SD	Min	Max
Seconds per meter (Avg.)	0.182	0.182	0.039	0.104	0.554
Seconds per meter (B-spline)	0.102	0.0986	0.051	0.0100	0.692
Taxi pickups	245.666	124	352.868	0	7198
311 complaints	0.038	0	0.256	0	13
311 complaint: blocked driveway (no access)	0.005	0	0.089	0	13
311 complaint: blocked driveway (partial access)	0.001	0	0.038	0	7
311 complaint: blocked roadway (construction)	0.002	0	0.044	0	4
311 complaint: failed roadway repair	0.004	0	0.074	0	6
311 complaint: pothole	0.024	0	0.194	0	12
311 complaint: rough roadway	0.002	0	0.048	0	4
Bikelane (any)	0.089	0	0.286	0	2
Bikelane: wide parking	0.020	0	0.139	0	1
Bikelane: protected	0.064	0	0.245	0	1
Bikelane: standard	0.001	0	0.024	0	1
Bikelane: signed route	0.005	0	0.068	0	1
Citibike station	0.000	0	0.008	0	1
Pothole outstanding	0.047	0	0.242	0	6
Zoning: EC-2/EC-3 districts	0.073	0	0.261	0	1

Note: Taxi pickups includes all pickups along an avenue, not only within-avenue trips. Bikelane (any) takes a value of 2 at three bikelane intersections.

fact that the data contains approximately half a million trips per day to isolate quasi-experimental "taxi race" conditions in which both a yellow and a boro taxi both picked up and dropped off a passenger within 10 meters of each other, as well as within a minute of each other. We find 234 such natural experiments in our area of northern Manhattan. We then regress travel time outcomes on a fixed effect per "taxi race" and a binary indicator for the boro taxi in the pair. Results are reported in Table 3.3. We find positive but non-significant coefficients in travel time levels, logs or ratios (i.e. seconds-per-meter). Since the magnitude of the estimated coefficient is a non-trivial 2.5% speed difference, we explore relaxing the threshold on pickup time difference from one minute apart to two, and up to 12, and find results that are statistically significant at the 5% level in a couple specifications in logs, as well as in levels, although not in seconds-per-meter. Alternative specifications intended to control for within-experiment variation in observables (such as a "pole position" effect, number of passengers or the residual variation in trip distance on the taxi meter) do not affect the results. We take this collectively as evidence that there may be a small difference in driving speeds for boro and yellow taxis.[8]

[8]A reason for this speed difference could be that a higher proportion of boro taxis are wheelchair-accessible vans, and these vehicles may accelerate more slowly than the average taxi. Alternatively, since daily rental rates (for medallion and car) are lower for boro than yellow taxis, it may be the case that drivers respond to the different opportunity costs of time, or that the activity of driving a boro taxi selects drivers with a lower subjective opportunity cost of time, i.e. slower drivers.

Table 3.3: "Taxi race" Quasi-Experiments: Speed Comparison between Yellow and Boro Taxis

	(12)	(11)	(10)	(9)	(8)	(7)	(6)	(5)	(4)	(3)	(2)	(1)
Seconds	7.763	7.980	6.656	6.260	6.857	6.613	8.512	10.88	14.09	16.19	12.22	10.86
	(3.633)	(3.655)	(3.580)	(3.737)	(3.982)	(4.284)	(4.611)	(4.674)	(5.258)	(6.154)	(7.094)	(6.812)
log($Seconds$)	0.0220	0.0237	0.0148	0.0133	0.0129	0.0106	0.0140	0.0256	0.0373	0.0446	0.0351	0.0252
	(0.0171)	(0.0171)	(0.0161)	(0.0169)	(0.0177)	(0.0193)	(0.0216)	(0.0168)	(0.0187)	(0.0218)	(0.0265)	(0.0245)
Seconds-per-meter	-0.0805	-0.0808	-0.319	-0.340	-0.372	-0.416	-0.508	0.00379	0.00542	0.00621	0.000126	0.00289
	(0.396)	(0.399)	(0.342)	(0.365)	(0.398)	(0.444)	(0.511)	(0.0221)	(0.0260)	(0.0315)	(0.0428)	(0.00686)
Race FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	1,280	1,270	1,220	1,146	1,050	940	816	704	598	494	362	234

Note: All specifications are "taxi race" regressions in which the only regressor is a binary indicator for the boro taxi in a taxi pair, and all between-variation is absorbed by a fixed effect per taxi race, as in a standard "twins" specification. Rows indicate the outcome variable. Columns indicate the number of minutes difference in the pickup time that are allowed for a pair of taxi trip records to be classified as a "taxi race," so sample size increases as this threshold is relaxed. Standard errors clustered at the "race" level. Results are unchanged if controlling for remaining within-experiment variation in observables, such as a "pole position" effect, number of passengers or the residual variation in trip distance on the taxi meter, in the two specifications where the outcome is in units of time rather than a rate of change.

We develop two alternative estimates of travel time to address the issues discussed in this section, but before developing these we discuss a brief pre-existing literature: to our knowledge, no previous paper has recovered street speeds from taxi GPS "probe" data in a manner that resolves the aggregation issues described above, or employed such estimates for the evaluation of urban transportation policy. Three prior papers in the computer science literature, however, have developed methods to reconstruct street speeds for the purposes of exploratory and descriptive analysis within a large urban area. Zhan et al. (2013), Santi et al. (2014) and Poco et al. (2015) are concerned with recovering speeds on street segments from the TLC trip data given that trip origins and destinations are located on the NYC road network and the path taken by the driver is unknown. The approach taken in these papers is to employ a routing algorithm to determine a set of latent, unobserved paths, and then allocate total trip travel times to weighted segments from the latent paths. Since our focus is precision within a specific area of interest, rather than exploratory analysis, we circumvent the path attribution problem by employing only the subset of trips whose origin and destination are located along the same avenue and in the direction of traffic. Conditional on a given latent path, Santi et al. (2014) and Poco et al. (2015) average travel times over trips crossing a segment and are therefore identical to our approach in this section, whereas Zhan et al. (2013) project trip times onto segment dummies and are therefore identical to the approach we describe in the next section, although their approach does not contain controls for observable trip characteristics, (e.g. driver type, or distance from GPS coordinates to grid network), or allow for non-linearity arising from fixed and variable components of travel time cost (e.g. due to acceleration).

3.3.2 Projection on Dummies to Recover a Travel Time Function

Our first approach to estimating marginal travel times at the street level while controlling for spatial and temporal aggregation bias consists of projecting travel times onto dummy variables for the intervals that are traversed by each individual trip, as well as observables such as taxi type. This approach is intended as an intermediate illustration for our next approach, so we present it in a summary manner.

We predefine a set of interval bins $b = \{\underline{b}, \ldots, \bar{b}\}$ of varying width over several city blocks, and construct a set of indicators variables m_{bi} that are turned off at a value of zero if a taxi trip does not intersect the interval and are turned on at a value equivalent to the width of an interval if the trip intersects it (this is equivalent to regressing travel time on dummies and appropriately rescaling coefficients after estimation). We also define sets of binary indicators that turn on if trip i starts (s_{bi}) or ends (e_{bi}) in a given interval. For each

month and uptown or downtown "run" along an avenue, we project the travel time T_i of trip i onto these indicators:

$$T_i = \sum_{b=\underline{b}}^{\bar{b}} \beta_{mb} m_{bi} + \sum_{b=\underline{b}+1}^{\bar{b}-1} \beta_{sb} s_{bi} + \sum_{b=\underline{b}}^{\bar{b}} \beta_{eb} e_{bi} + \beta x_i + \delta_h + \varepsilon_i \tag{3.1}$$

where additional controls include hour fixed effects, two alternative controls for boro taxi status (a dummy, consistent for a fixed cost, and a trip distance measure interacted with boro taxi status, consistent for a boro variable cost) and a control for "surplus distance" on the taxi meter after the coordinates of each trip endpoint undergo a perpendicular projection onto the straight north-south line along each avenue. Each trip will also turn on an indicator dummy both for both endpoints, as well as a number of middle indicators. Because we observe trips that fully traverse the intervals $b = \{\underline{b}+1, \ldots, \bar{b}-1\}$, and have saturated the specification with terms that can absorb contributions to travel time from trip events other than the crossing of these intervals, the estimated coefficients $\hat{\beta}_{mb}$ for $b = \{\underline{b}+1, \ldots, \bar{b}-1\}$ are consistent for (an average over) the interval marginal second per meter function specific to bin b, whereas the coefficients on $\hat{\beta}_{mb}$ for $b = \{\underline{b}, \bar{b}\}$ are consistent for an average of the marginal second per meter function over the fractional measures of the interval that are intersected in the data, which are not a quantity of interest and are discarded. All of the coefficients on $\hat{\beta}_{sb}$ and $\hat{\beta}_{eb}$ play a similar role, as they are consistent for average marginal costs on fractionally traversed intervals, and are also discarded. Given that trip data is only observed in the interior of the interval defined by the lower limit of $\underline{b}$ and the upper limit of $\bar{b}$, the average speed parameters β_{mb} for $b = \{\underline{b}, \bar{b}\}$ are not identified. The estimated coefficients $\hat{\beta}_{mb}$ for $b = \{\underline{b}+1, \ldots, \bar{b}-1\}$ constitute our second measure of an underlying "seconds-per-meter" travel time function. These estimates control for spatial aggregation bias, since travel times are projected linearly onto the event of crossing specific intervals, and so estimates are computed from differences in the overlaps between intervals traversed. Continuing with the hotel example from above, if variation elsewhere in the data identifies downtown segments as being particularly slow, this high time cost will be subtracted from the southbound trips originating at the hotel, allowing travel time in the hotel's interval to be identified from that interval's actual contribution to travel time (lower, in our example). This method also allows us to include hour fixed effects to control for the time aggregation bias that would arise from correlation between trip endpoints and time of day, as well as controls for trip observables.

The projection on dummies method does have some costs, however, as it involves

substituting variance for bias. Discretizing the "second per meter" travel time function into a set of intervals leads to loss of information, as the method discards average travel time data on the portion of each trip that does not fully traverse an interval. Discretization cannot be made arbitrarily low in a finite sample: it will eventually lead to intervals that do not originate or end trips, causing identification failure as the projection matrix loses full rank. Since estimates in the projection method are obtained from differencing time costs across intervals, errors in one interval can propagate to others, leading to high variance of estimates. The discontinuity in the seconds-per-meter function that is imposed as an assumption on the data generating process is one source of variance, as it leads to abrupt changes in the travel times that are attributed to otherwise similar trips if only one of these crosses the junction point used in discretization.[9] To address these concerns, our next method also controls for potential aggregation bias while delivering a better trade-off in terms of reduced variance.

3.3.3 Projection on B-splines to Recover a Travel Time Function

Our second projection method controls for aggregation bias and observable trip characteristics in the same way as the projection dummy method above, but explicitly models a continuous "seconds per meter" function with basis splines: the advantages of this method are that it eliminates discontinuities at junction points between interval dummies, and that it eliminates data loss on trip segments that do not fully traverse a discrete interval, both of which reduce the variance of estimates. Additionally, this method produces a flexible estimate for the "seconds per meter" function that can be evaluated at every point of its support.

Let s be the true "seconds per meter" function, which is defined over an interval of road during a given month. A taxi trip i that begins at point a_i and ends at point b_i on the interval will take a total time:

$$T_i = \beta_1 x_{1i} + \int_{a_i}^{b_i} \left[s(\tau) + \beta_2 x_{2i} \right] d\tau + \varepsilon_i \tag{3.2a}$$

$$T_i = \beta_1 x_{1i} + S(b_i) - S(a_i) + \beta_2 (b_i - a_i) x_{2i} + \varepsilon_i \tag{3.2b}$$

where x_{1i} are trip characteristics that do not depend on location or trip distance (i.e. fixed time costs), x_{2i} are trip characteristics that depend on trip distance but not location (variable time costs) and $S(b_i) - S(a_i)$ is the area under the seconds-per-meter function.

[9] In our application, we employ as few at seven intervals and obtain estimated travel times that are *negative* in 1.7% of all run-bin-month cells.

We can estimate (Eq. 3.2b) by regressing T_i on the spline differences $[S(b_i) - S(a_i)]$, x_{1i} and $(b_i - a_i)\, x_{2i}$, which we can implement using basis splines for the location of trip origins and destinations, provided we impose the constraint that the spline function is the same at both endpoints.[10]

To implement B-splines, if we let:

$$S(\tau) = \sum_{k=1}^{K} \gamma_k \psi_k(\tau)$$

where ψ_k are the bases of the spline, then the regression above can be written as:

$$T_i = \beta_1 x_{1i} + \sum_{k=1}^{K} \gamma_k \left[\psi_k(b_i) - \psi_k(a_i)\right] + \beta_2 (b_i - a_i)\, x_{2i} + \varepsilon_i \tag{3.3}$$

which can be estimated by OLS. Since the bases $\psi_k(\tau)$ are differentiable, after estimation we can recover

$$\hat{s}(\tau) = \hat{S}'(\tau) = \sum_{k=1}^{K} \hat{\gamma}_k \psi_k'(\tau)$$

where $\hat{\gamma}_k$ are coefficients estimated by OLS.

3.4 Impact of the Boro Program on Traffic Congestion

3.4.1 Empirical Strategy

Historical data on street speed at a high level of spatial resolution allows us to evaluate the impact of policies that impact road use in a well-defined spatial manner. In this

[10]Our specification for the error term assumes that it is additive to integrated travel time. Alternatively, we could assume an instantaneous mean zero error that arises over the interval. In this case we would have that $T_i = x_{1i} + \int_{a_i}^{b_i} s(\tau) + \beta x_{2i} + \varepsilon_i(\tau)\, d\tau$ and so we could write $T_i = x_{1i} + S(b_i) - S(a_i) + \beta (b_i - a_i)\, x_{2i} + (b_i - a_i)\, \bar{\varepsilon}_i$ where $\bar{\varepsilon}_i = (b_i - a_i)^{-1} \int_{a_i}^{b_i} \varepsilon_i(\tau)\, d\tau$ is a heteroskedastic error term. We could estimate this equation by GLS, or divide both sides by $(b_i - a_i)$:

$$\bar{s}_i = \frac{T_i}{(b_i - a_i)} = \beta_1 \frac{x_{1i}}{(b_i - a_i)} + \frac{S(b_i) - S(a_i)}{(b_i - a_i)} + \beta_2 x_{2i} + \bar{\varepsilon}_i$$

where $\bar{s}_i = \frac{T_i}{(b_i - a_i)}$ is a trip's average seconds per meter. We can then implement a regression of $\bar{s}i$ on rescaled splines differences $\left[\frac{S(b_i) - S(a_i)}{(b_i - a_i)}\right]$ and x_i, which imposes the GLS correction for heteroskedasticity under the assumption of integrated errors.

section we describe how we exploit the roll-out of the Street-Hail Livery program to evaluate the impact of boro taxis on congestion in northern Manhattan. Our area of study spans 1 kilometer of the Upper East Side and 2 km of East Harlem on the East Side, and 2 km of Upper West Side and 1 km of Morningside Heights and Harlem on the West Side. Urban density and socioeconomic status decline towards the north, although not in a monotone or smooth manner. Prior to the launch of the boro program, the declining gradient in urban density was associated with lower yellow taxi pickup activity, with spikes around major intersections, which include subway stations.[11] The boro taxi program was intended to increase supply in areas undeserved by traditional yellow taxis, but its implementation through a "hail-exclusion" zone south of East 96th and West 110th streets creates an incentive for boro taxi drivers to cruise for hails in the area just north of the boundary: whereas this area was of marginal cruising value for yellow taxi drivers, it is a corner-solution for boro taxi drivers, leading to a localized shock to supply.[12]

Our empirical strategy in this section consists of exploiting this localized shock, and its roll-out over time, to evaluate the impact of boro taxis. Figures 3.2 and 3.3 show the average travel time measures (estimated as described in Section 3.3.1) for all the avenues that contain useful variation to identify the main results in the paper.[13] For ease of presentation the data has been averaged at the annual level, and expressed in miles per hour, rather than seconds per meter. Inspection of these figures reveals the identifying variation in the data and that drives our main difference-in-differences estimates: historical speeds increase in the northern direction along most avenues, along with declining urban density and traffic,[14] and as time goes on all streets become slower, but particularly so north of the hail-exclusion boundary, an area we refer to as "boro-zone." Particularly stark examples include 5th Ave (Fig. 3.3a) and Lexington Ave (Fig. 3.3e), but the result can be observed on most avenues.

We employ a difference-in-differences specification to estimate an average reduced-form impact from the boro taxi program on street speed in northern Manhattan. Let s_{rbm}

[11] As shown in Fig. 3.12 most yellow taxi activity is concentrated at airports as well as midtown and downtown Manhattan.

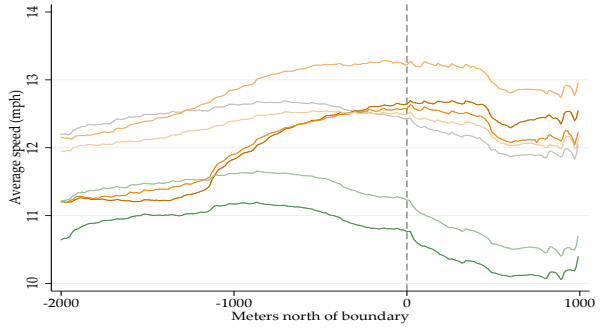
[12] This notion holds under a spatial search equilibrium as in Buchholz (2017): areas with high baseline demand support higher search in equilibrium, although the endogenous continuation value for above-boundary pickups may be shaded down by a higher transition probability into the hail-exclusion zone.

[13] Some avenues, such as West End, Amsterdam downtown or Columbus uptown do not cross the hail-exclusion boundary and therefore do not contain identifying variation after we include fixed effects at the avenue level. We also drop Manhattan Ave since it does not contain sufficient trips north of the boundary to reliably estimate speed.

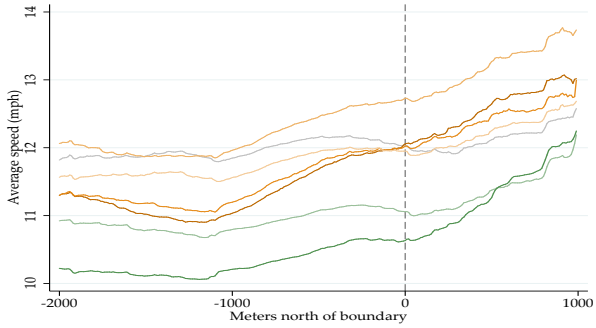
[14] This is always the case on the East side, but not on the West side where some avenues both narrow and become two-way north of the hail exclusion boundary. The time-invariant attributes of streets, such as changes in width, the direction of traffic, or bends in the roadway will be absorbed in specifications described below by fixed effects at the level of the avenue run and 10 meter bin.

Figure 3.2: Average Travel Time at the Avenue Level, Over 10 Meter Intervals. West Side

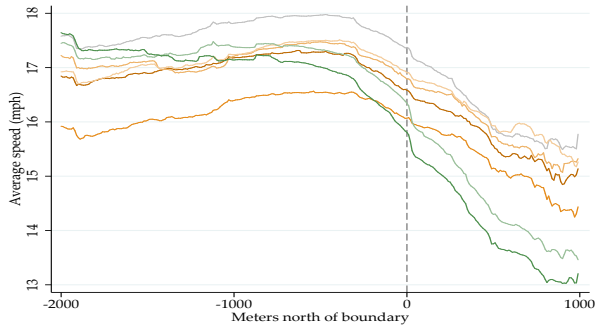
(a) Broadway Northbound



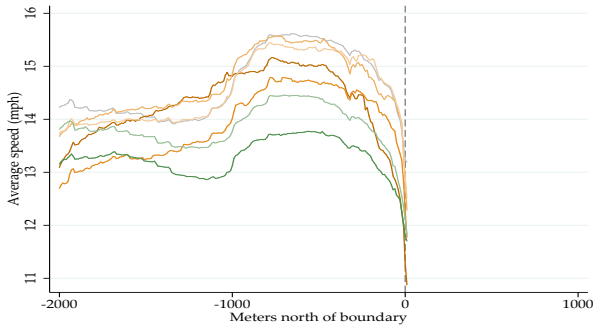
(b) Broadway Southbound



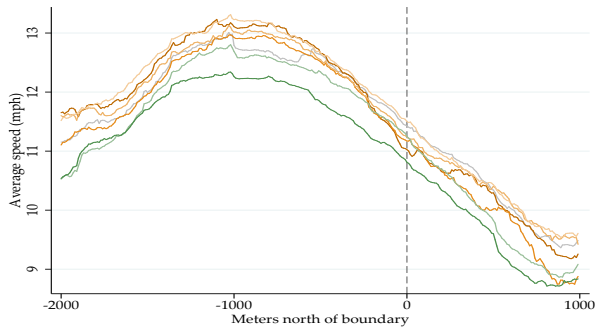
(c) Amsterdam Northbound



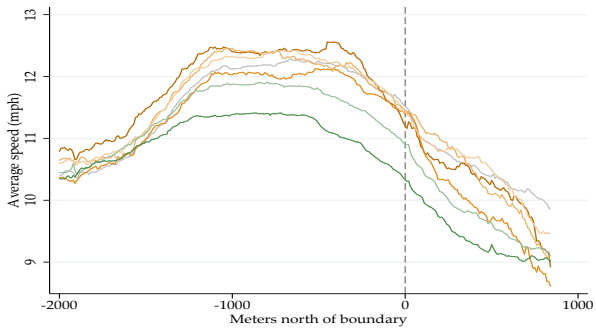
(d) Columbus Southbound



(e) Central Park West Northbound



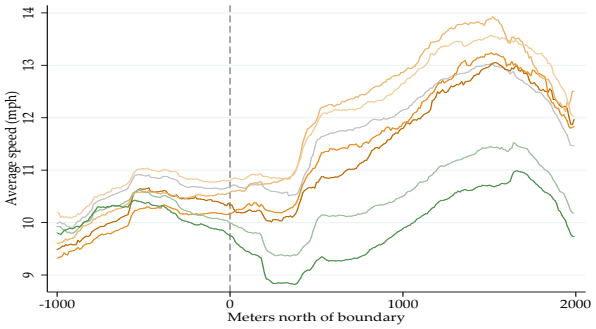
(f) Central Park West Southbound



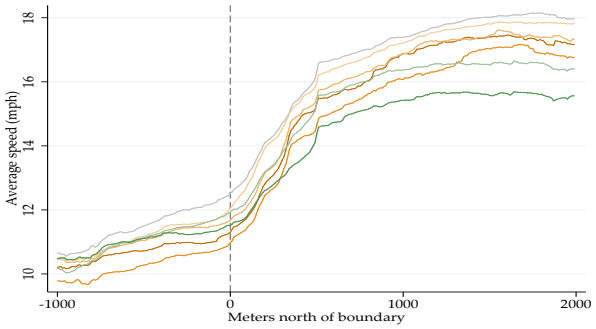
Notes: Average travel time is computed as the average seconds per meter at the 10 meter interval over all trips that cross the interval. Averages shown here are annual and have been converted to miles per hour. Year is indicated by color and shade, according to the following scheme: years prior to the boro taxi program are indicated in orange, from 2009 (darkest) to 2012 (lightest). 2013 is shown in gray, as program roll-out begins on Aug 8, 2013. 2014 is shown in light green, and 2015 in dark green. 2016 not shown to control for seasonality, since data is only available for the first half of the year. Not plotted: both directions of West End Ave, which are included in regressions but do not identify the coefficient of interest because, like Columbus Ave, they do not cross the hail-exclusion boundary. Manhattan Ave is excluded from plots and regressions because in its short length it does not generate sufficient within-avenue trips to allow for speed measurement.

Figure 3.3: Average Travel Time at the Avenue Level (continued). East Side

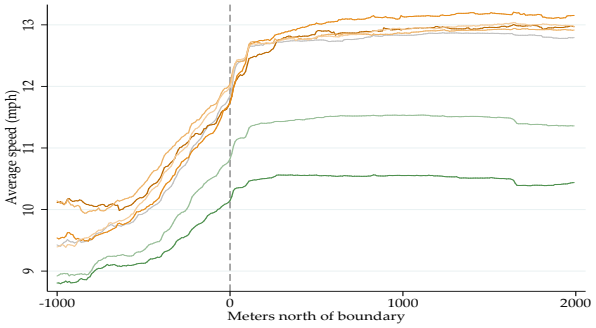
(a) 5th Avenue Southbound



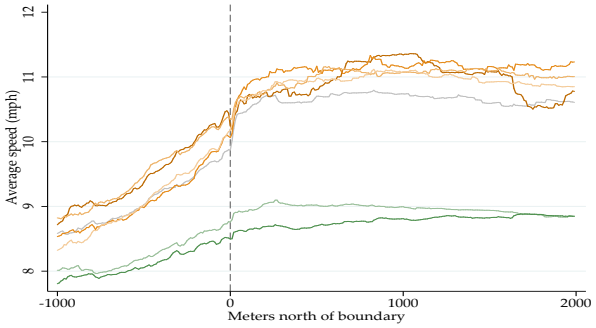
(b) Madison Avenue Northbound



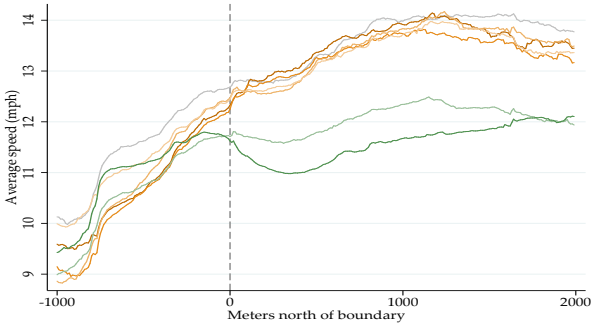
(c) Park Avenue Northbound



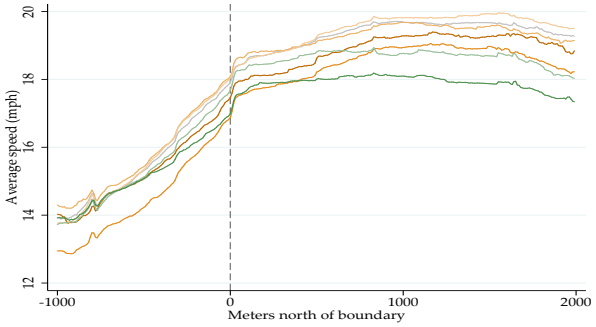
(d) Park Avenue Southbound



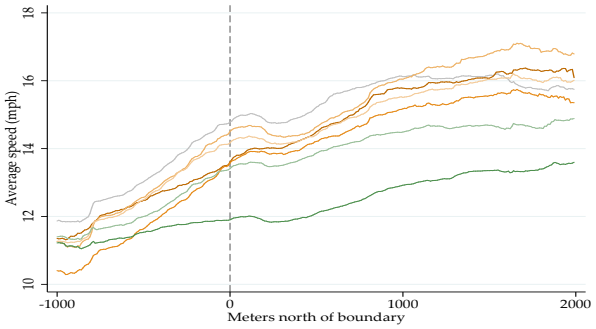
(e) Lexington Avenue Southbound



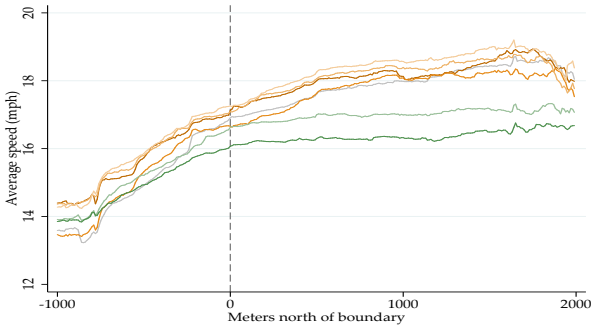
(f) 3rd Avenue Northbound



(g) 2nd Avenue Southbound



(h) 1st Avenue Northbound



be a travel time outcome, which in most specifications is the log seconds per meter at the level of an avenue run (r), 10 meter bin (b) and month (m) from January 2009 to June 2016. In a standard difference-in-differences specification we would define the post dummy as August 2013 and all months thereafter:

$$s_{rbm} = \beta_{DID}\text{post}_m \times \text{boro zone}_{rb} + \beta x_{rbm} + \text{fixed effects} + \varepsilon_{rbm} \tag{3.4}$$

However, the supply of boro taxis was rolled-out gradually: only 114 boro taxis carried out at least one trip during the program's first month of August 2013, and this number rose to 445 taxis in September, 1,128 in October, 1,754 in November and 2,736 in December. Although this ramp-up was relatively fast, it took until June 2015 for the program to reach the highest number of boro taxis ever observed to be active in a given month, which was 6,539. The time series for the number of unique boro taxis is plotted as the solid green line in Figure 3.4. We define the $boro\,roll\text{-}out_m$ treatment intensity measure as the ratio of boro taxis active each month (zero prior to August 2013) to the maximum value of 6,539, and our preferred specification is to employ this intensity measure in difference-in-differences specifications of the form:

$$s_{rbm} = \beta_{DID}\text{boro roll-out}_m \times \text{boro zone}_{rb} + \beta x_{rbm} + \text{fixed effects} + \varepsilon_{rbm} \tag{3.5}$$

Our baseline specifications include fixed effects δ_{rb} at the level of each run and bin, absorbing time invariant features of the road, as well as a set of monthly time-effects δ_{rm} at the level of each run, which absorb the variation in speed along each avenue over time that is common across bins. The main difference-in-differences coefficient under this fixed effect specification is estimated from the differences within each avenue between the travel time north vs. south of the hail-exclusion boundary as the boro program was rolled out. For regressions based on eqs. (3.4) or (3.5) to provide an estimate of the causal impact of the boro program we require the parallel trend assumption that street speed in the hail-exclusion "control" zone and boro "treatment" zone are not trending apart from each other for reasons other than the impact of the program. We will discuss below the robustness of our estimates to trends in unobservables, along with alternative fixed-effect structures and specifications, but first we consider a number of observable features of the urban environment that vary at the level of road segments over time and could act as confounders for our estimate of the boro program's impact on traffic speed.

A first source of potential, observable confounders consists of changes in the use

[15]See Sadik-Khan and Solomonow (2016) for the former NYC Department of Transportation commissioner's account of the substantial expansions in bike lanes and pedestrian plazas over this period.

of road space, for example for transportation alternatives. Bike lanes were expanded substantially throughout NYC during the period of our study, so we use time-stamped maps of the bike lane network to control for bike lane expansion in our area of study.[15] A second change in transportation infrastructure with a potential impact on road use and traffic speed was the deployment and northward expansion of the Citibike bikeshare system. Citibike launched south of Central Park - outside our area of interest - in May 2013, but expanded into the upper east and west sides over the summer of 2015. We map the deployment of Citibike stations over time as well as, in an alternative specification, all areas along an avenue south of any newly opened Citibike station, to control for both stations and cyclists.[16] Additional potential confounders include changes to road conditions. We construct count measures at the road-segment level for complaints reported to the city's "311" municipal services hotline, for conditions that could impinge traffic. We also construct count measures for outstanding potholes. Additionally, two re-zoning events took place during our period of study, the establishment in June 2012 of Special Enhanced Commercial Districts 2 and 3 along a number of avenues on the Upper West Side. Although this re-zoning was entirely commercial in character, intended to regulate "ground floor frontages" to preserve commercial activity and "multi-store character," in some specifications we include a control for these zoning changes. Details on the construction of these controls are in 3.8.1.

3.4.2 Results

We report the results for our baseline specifications on Table 3.4. The table is organized to report results for our three alternative measures of travel time: average speeds as in Section 3.3.1 in Cols. 1 and 2, B-spline projected speeds as in Section 3.3.3 in Cols. 3 and 4, and dummy projected speeds as in Section 3.3.2 in Cols. 5 and 6. Within each of these, we report the results for Eqs. (3.4) and (3.5).

[16]See Molnar and Ratsimbazafy (2016) for a study of the Citibike bikeshare program and its impact on the medallion taxi industry. Hamilton and Wichman (2017) find the bikeshare infrastructure reduced traffic in Washington D.C.

Table 3.4: Reduced Form Impact of Boro Program on Congestion.
Comparison Across Travel Time Estimation Methodologies

	(1) Log avg. s/m (Treat dummy)	(2) Log avg. s/m (Baseline)	(3) Log Bsp. s/m (Treat dummy)	(4) Log Bsp. s/m (Baseline)	(5) Log est. s/m (Treat dummy)	(6) Log est. s/m (Baseline)
BZ x Post	0.063		0.069		0.062	
	(0.004)		(0.017)		(0.037)	
BZ x Roll-out		0.081		0.101		0.083
		(0.004)		(0.020)		(0.045)
Bikelane	0.026	0.025	-0.091	-0.094	0.084	0.084
	(0.004)	(0.004)	(0.053)	(0.053)	(0.036)	(0.036)
Citibike station	-0.018	-0.009	0.089	0.103	-0.085	-0.071
	(0.007)	(0.007)	(0.036)	(0.037)	(0.012)	(0.015)
311 complaints	-0.000	-0.000	0.003	0.003	-0.003	-0.003
	(0.000)	(0.000)	(0.003)	(0.003)	(0.003)	(0.003)
Potholes	-0.001	-0.001	-0.006	-0.006	-0.008	-0.008
	(0.001)	(0.001)	(0.009)	(0.009)	(0.008)	(0.008)
Run-bin × Run-my FE	Y	Y	Y	Y	Y	Y
Observations	406260	406260	406260	406260	6849	6849
R^2	0.97	0.97	0.64	0.64	0.53	0.53
R^2-within	0.13	0.15	0.0015	0.0022	0.0052	0.0058

Note: Unless otherwise specified, robust standard errors are two-way clustered at the level of run× bin and run× month. The difference in difference coefficient in the baseline regression (Col. 2, row 2) remains significant at 0.1% with coarser clustering schemes, such as two-way clustering at the level of runs and months (s.e. 0.0151), along any single dimension, or under alternative schemes (and bandwidths) with clustering on the time or time-run dimensions with Newey-West robustness to autocorrelation along the spatial dimension by bin or run-bin. Further evaluation of inference is provided in Table 3.5, Cols 5 & 6.

Comparing either Cols. 1, 3 and 5 or 2, 4 and 6 we can see that results for a given specification are similar across alternative estimation methods for the travel time outcome.[17] In all specifications it is also the case that the DID coefficient estimate on the $boro \times post$ dummy is lower than on the $boro \times roll\text{-}out$ intensity measure, which is what we would expect if a partially-rolled out boro program has a less than full impact on traffic speed. Once the full roll-out of the program is allowed for, we find travel time increases of 8-10% in the boro zone relative to the hail-exclusion zone.[18] To further evaluate the role of timing and pre-trends in our specification, we re-run a version of Eq. (3.5) in Col. 2 in which the pre-treatment period is defined to start 24 months prior to the actual introduction of the boro program (i.e in August 2011) and we estimate a month-specific $\hat{\beta}_{DID,m}$ coefficient for every month thereafter. Figure 3.4 plots (in blue) the point estimates and 95% confidence intervals for these coefficients: we can see that there are no pre-trends, as monthly coefficients are flat at zero prior to the actual roll-out of the program. The speed difference between the hail-exclusion and boro zones starts to develop shortly after August 2013, quickly trending upward over the first six months, and then more slowly over the next year. The trend in $\hat{\beta}_{DID,m}$ coefficients tracks the trend in the roll-out of boro taxis: the measure of this roll-out is not used in the estimates plotted in Figure 3.4, but it is overlaid on the figure for comparison.[19]

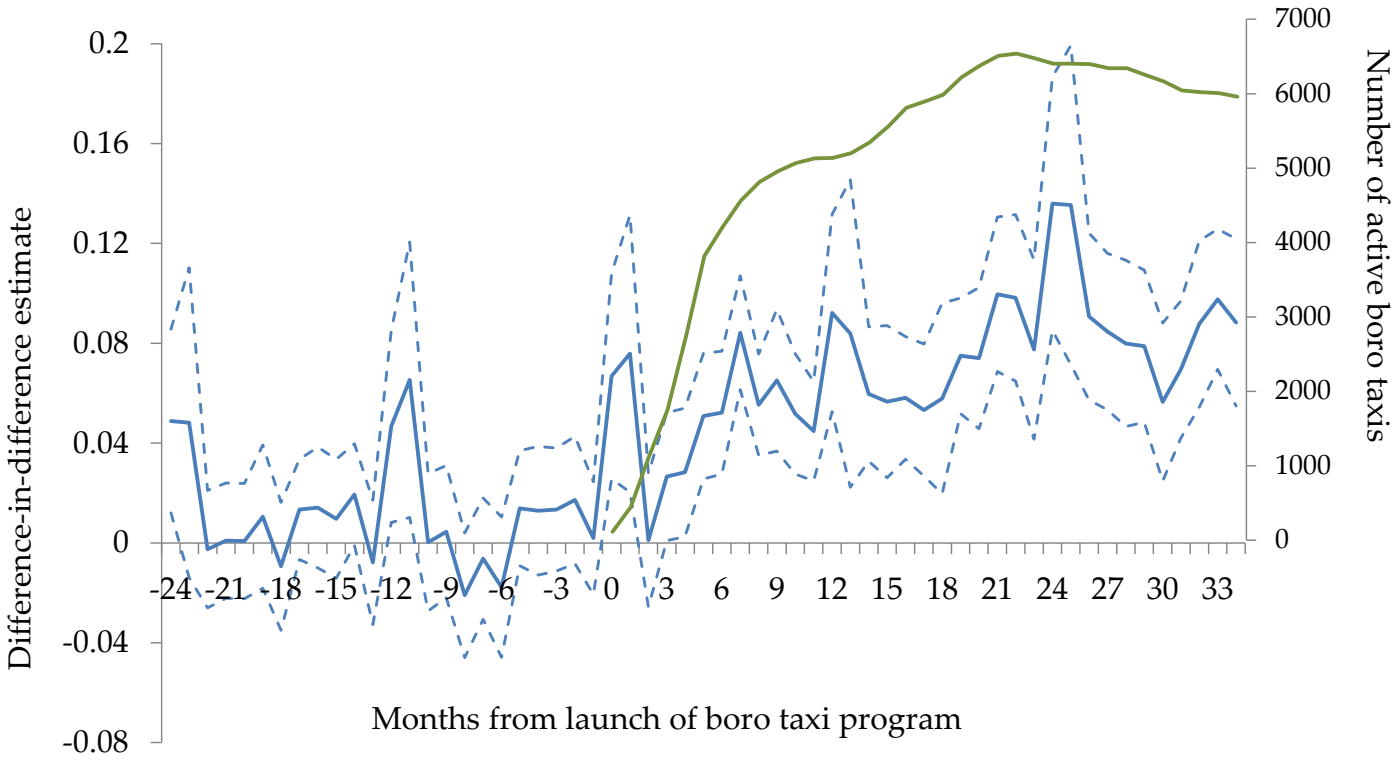
In Table 3.5 we evaluate some alternative specifications to assess our claim that we have estimated an 8-10% causal impact of the boro program on traffic speed (estimates of $\hat{\beta}_{DID}$ can be read equivalently as log increases in travel time or log decreases in speed). Col. 1 presents an estimate in levels, rather than logs, which is statistically significant and consistent with our baseline results: at the sample average of 0.182 seconds per meter (i.e. 12.32 miles per hour) this estimate represents a 7.2% decrease in speed (to 11.44 miles per hour). Col. 2 presents estimates with common time effects (i.e. fully saturated at the run-bin and at the monthly level, but not interacted with runs). Relative to our baseline,

[17] We employ an exhaustive set of fixed effects: this results in an R^2 well above 0.9 for specifications (1) and (2), but declines to around 0.5-0.6 for projection-based travel times due to higher variance in that outcome variable.

[18] If we had defined "full roll-out" as a ratio relative to a lower denominator (e.g. the average number of boro taxis on the road after a certain period, rather than the maximum of 6,539 taxis) then the estimated impact from the program on traffic speed would be slightly higher, but there would be no effect on the congestion estimates reported in Section 3.5 below.

[19] Close attention to Figure 3.4 reveals that the "dynamic DID" coefficient estimate spikes in August of each year. These spikes are explained by a seasonal violation of the parallel trends assumption in our DID specification: every August (and to a lesser extent, September) there is a vacation-driven decline in both the population of taxi users and drivers in NYC (see Figure 3.1 for a monthly time-series of taxi trips that displays negative seasonality in August). The vacation-induced reduction in traffic and increase in speed is stronger south of the hail-exclusion boundary, resulting in the seasonal variation in travel time differences displayed in Figure 3.4.

Figure 3.4: Difference-in-Differences Estimates of the Impact of the Boro Program, by month



Notes: In green, the number of unique boro taxis active every month, scale on the right axis. In blue, difference-in-differences coefficients and 95% point-wise confidence intervals for standard errors clustered at the run-bin level. The baseline period is defined as the 31 months from Jan 2009 to Jul 2011, thus ending two years prior to the launch of the boro program. A separate coefficient is estimated for every month thereafter. Months are labelled on the plot relative to August 2013. Coefficient estimates are from Eq. (3.5), and include a fixed effect at the level of the run, 10m bin and month of the year to absorb bin-specific seasonality.

this specification allows variation in travel time between runs over time to identify the difference-in-difference estimates.

Table 3.5: Reduced Form Impact of Boro Program on Congestion.
Alternative Specifications

	(1) Avg. s/m (Levels)	(2) Log avg. s/m (Rect. FE)	(3) Log avg. s/m (Linear trend)	(4) Log avg. s/m (Lin. & rideshare trend)	(5) Log avg. s/m (Aggr. T)	(6) Log avg. s/m (Aggr. T,Run)
BZ x Roll-out	0.014	0.064	0.067	0.062		
	(0.001)	(0.005)	(0.007)	(0.007)		
Bikelane	0.003	0.023	0.025	0.024	0.016	
	(0.001)	(0.005)	(0.004)	(0.056)	(0.025)	
Citibike station	-0.001	-0.029	-0.016	-0.008	-0.104	
	(0.001)	(0.020)	(0.007)	(0.084)	(0.106)	
311 complaints	-0.000	0.001	-0.000	-0.000	0.012	
	(0.000)	(0.001)	(0.000)	(0.038)	(0.014)	
Potholes	-0.000	0.001	-0.000	-0.000	0.030	
	(0.000)	(0.002)	(0.001)	(0.107)	(0.065)	
BZ x Post					0.050	0.044
					(0.017)	(0.016)
Post					0.057	0.058
					(0.012)	(0.012)
BZ dummy						-0.100
						(0.046)
Run-bin × Run-my FE	Y		Y	Y		
Run-bin × my FE		Y				
Run × BZ linear trend			Y			
Run × BZ rideshare trend			Y	Y		
Observations	406260	406260	406260	406260	9028	60
R^2	0.96	0.93	0.97	0.98	0.98	0.85
R^2-within	0.099	0.055	0.20	0.22	0.73	0.34

Note: Unless otherwise stated, robust standard errors are two-way clustered at the level of run× bin and run× month. Estimates with linear and "aggregate ridehail" trends are interacted with the treatment zone and run, but results barely change if interacted at the treatment level only. Estimates in Col. 5 include fixed effects and are clustered at the level of run× bin and run× post, likewise by run× boro-zone and run× post for estimates in Col. 6.

The major threat to a causal interpretation of the estimates from our research design is that unobservable components of traffic speed could have different trends in the hail-exclusion "control" and boro "treatment" zones, for reasons other than the program's impact. We check against any such linear trends in Col. 3, and find that the estimated impact decreases by only 17% after the inclusion of a linear trend interacted by run and treatment status. Robustness to a difference in unobserved linear trends is a strong test for the validity of a difference-in-differences design (since identification is provided by the timing of the boro program's launch in August 2013 and its subsequent roll-out relative to any linear trend), although this is less the case in a setting such as ours in which all treated units change status at the same time.

We cannot check for robustness against arbitrary, non-linear differences in trends between the treatment and control zones, since the set of these trends contains the roll-out of the boro program that we rely on for identification. We can, however, check the robustness of our estimates against an alternative, well-defined trend that may be a cause for concern. The major recent change to NYC transportation that we have not yet addressed in detail was the growth in ridehail (plotted in Fig. 3.1), starting with Uber's entry as a black-car service in May 2011. If growth in ridehail played a role in the substantial slowdown of NYC's streets since 2013 (as documented in 3.8.3, and for our area of interest by the downward, parallel translation over time of the speed curves shown in Figures 3.2 and 3.3) then it could bias our estimates: although our research design controls for ridehail's potential effect on traffic speed to the extent that this effect was proportional both north and south of the hail-exclusion boundary, differential impacts across zones remain a potential confounder. To address this concern we construct an "aggregate ridehail trend" from monthly ridehail trips, and interact this trend with runs and the treatment zone.[20] In Col. 4 we jointly include the ridehail trend along with a linear trend: we find that our reduced-form estimate of the impact of the boro program is robust to a trending differential unobservable that is not only linear, but is allowed to accelerate after May 2011 at any rate proportional to the growth of ridehail in New York City.

We present Cols 5 and 6 as checks on statistical inference.[21] In Col 5 we present the results from averaging, for each spatial unit, all time periods into a before and an after

[20]Trip-level reporting by ridehail providers to the TLC started with the second and third quarters of 2014. There is a gap in reporting during the fourth quarter of 2014, and reporting has resumed from 2015 onwards. We interpolate the monthly growth rate for the fourth quarter of 2014, and extrapolate backwards to Uber's launch in May 2011 using the 10.1% monthly growth rate that has been observed for all ridehail providers since the second quarter of 2014.

[21]Standard errors in all other specifications are two-way clustered at the level of run× bin and run× month. Statistical significance holds at coarser clusters such as runs.

period. This is clearly not a preferred specification, as it eliminates variation in observables over time (from controls, such as bike lanes, or the identification provided by the gradual roll-out of boro-taxis that was documented in Fig. 3.4), but it is useful as a check that statistical significance is not an artifact of time series dependence. We then collapse all variation between road intervals into a single segment average for the treatment or control zone, as a further check that inference does not depend on unmodeled spatial dependence along each avenue. Col 6 presents the difference-in-differences estimate from this further step, which averages each avenue into four cells: both time periods vs. the avenue's treatment and control segments. Under this specification, which contains only 62 observations (and a fixed effect for each of 17 avenue runs, of which 16 are included in the control zone and 15 in the treatment zone) we still find a statistically significant 3.8% additional slowdown of the boro zone relative to the hail-exclusion zone control, which in turn slowed down by 5.8%.[22]

In Table 3.6 we present results under alternative data samples. Cols 1 and 2 break out results into the east and west sides of northern Manhattan: we find a heterogeneous impact of the boro program, with double the travel time increases on the east than on the west side. As we will describe below, this result is associated to a larger increase in taxi activity on the east side. Cols 3 and 4 look at narrow event windows around the introduction of boro taxis (6 and 12 months before and after, respectively). These results confirm that the boro program's impact is identified from variation in traffic speed around the roll-out period (as shown previously by the month-level "dynamic difference-in-differences" results plotted in Fig. 3.4), rather than long-run trends or events outside the period in which the program was introduced. In particular, results for both windows exclude a city wide-reduction in the speed limit from 30 mph to 25 mph, which took place on November 7, 2014.[23] Speed limits were changed along the entire length of the avenues in our study, so our baseline specifications already control for avenue-level effects through the monthly time-effects δ_{rm} at the level of each run. Results from Cols. 3 and 4 further confirm that the speed limit change does not introduce a violation of the parallel-trends assumption.

[22]We omit averaged controls, since at this level of aggregation they become an arbitrary set of free parameters. Our estimate of the $\hat{\beta}_{DID}$ coefficient is still significant and higher in magnitude if we include averages of our controls, or if we restrict our sample to the balanced panel of 14 runs that span both the treatment and control zones. Results employing the B-spline projection travel time outcome are consistent in sign and magnitude, but not significant at this level of aggregation.

[23]The speed limit change was part of the broader "Vision Zero" traffic safety program, announced in January 2014 by NYC mayor Bill de Blasio and rolled-out in stages thereafter. The program consists of a broad set of initiatives intended to eradicate traffic fatalities, including public service announcements, a focus on enforcement of traffic laws, street intersection redesign and the change in speed limits. See www1.nyc.gov/site/visionzero/initiatives/initiatives_archive.page for further details.

Table 3.6: Reduced Form Impact of Boro Program on Congestion.
Alternative Samples

	(1) Log avg. s/m (East)	(2) Log avg. s/m (West)	(3) Log avg. s/m (6m pre/post)	(4) Log avg. s/m (12m pre/post)	(5) Log avg. s/m (cutout 250m)	(6) Log avg. s/m (cutout 500m)	(7) Log avg. s/m (cutout 750m)	(8) Log avg. s/m (DiDiD)
BZ x Roll-out	0.100	0.049	0.042	0.077	0.086	0.090	0.086	
	(0.006)	(0.005)	(0.013)	(0.008)	(0.005)	(0.005)	(0.005)	
BZ x Roll-out x Uptown								-0.030
								(0.006)
Bikelane	0.017	0.027	0.018	0.043	0.026	0.026	0.025	-0.014
	(0.006)	(0.007)	(0.013)	(0.009)	(0.004)	(0.004)	(0.004)	(0.007)
Citibike station	-0.002	-0.012			-0.005	-0.001	-0.001	-0.019
	(0.010)	(0.003)			(0.007)	(0.007)	(0.008)	(0.007)
311 complaints	-0.000	-0.000	0.000	-0.000	-0.000	-0.000	0.000	-0.000
	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Potholes	-0.002	0.002	-0.001	-0.001	-0.000	-0.000	0.000	-0.001
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Run-bin × Run-my FE	Y	Y	Y	Y	Y	Y	Y	Y
BZ-my FE								Y
Observations	216000	190260	58682	112850	374760	339840	303840	200430
R^2	0.97	0.97	0.98	0.98	0.97	0.97	0.98	0.98
R^2-within	0.21	0.072	0.028	0.15	0.16	0.16	0.13	0.0098

Note: Unless otherwise specified, robust standard errors are two-way clustered at the level of run× bin and run× month. Standard errors for the triple-difference estimation in Col. 8 are additionally clustered at the level of boro-zone × month.

The specifications in Cols. 5-7 are presented as a check for spatial spillovers over the treatment boundaries. Some spatial trends are in fact present in the speed data: inspection of Figures 3.2 and 3.3 (see, in particular, the northbound direction of Broadway in Fig. 3.2a, as well as 3.3a and 3.3g) shows that following the program's roll-out, the speed on some avenues starts to decline in areas neighboring the hail-exclusion to the south. Unlike pre-trends in time, which we would not expect the program to produce, it is reasonable to expect the boro program to produce some form of spatial trend due to spatial spillovers. The channels though which the boro program could produce endogenous spillovers include: i) the direct effect of boro taxis on traffic as they drive south of the boundary to drop off passengers; and the incidence of such travel should diminish with distance to the boundary, ii) the direct effect of yellow taxis (or ridehail providers), which may retrench south of the boundary in response to the increased, localized competition from boro taxis to the north, and iii) an indirect effect transmitted through traffic congestion: a taxi may not only have a direct effect that slows down traffic at its immediate location, but also down the road through the effect of queuing cars, or on adjacent roads through gridlock (Schwartz (2015)).

Spatial trends of this form need not constitute a threat to the detection of an effect from the program, but rather to whether our difference-in-difference strategy identifies the program's average effect. In each of the cases outlined above, a spatial spillover from the program slows down traffic in the area we use as a control, biasing our coefficient estimate downward from the program's average effect. To evaluate the sensitivity of our results to spillovers, in Cols. 5-7 of Table 3.6 we drop the intervals in the control zone that are 250, 500 or 750 meters immediately south of the boundary. The estimated effect rises as we drop adjacent intervals, and we estimate an effect of 0.089 (i.e. 11% higher) if we exclude the 500m adjacent to the treatment boundary, but the sensitivity to this exclusion does not change after this point.

In Col. 8 we employ an alternative specification to directly estimate the third channel for spillovers, i.e. knock-on effects through queuing and road congestion. The degree to which the shock to activity in the boro zone can transmit traffic congestion across the treatment boundary will depend on the direction of traffic: avenues in the downtown direction (i.e. southbound) cannot, by definition, transmit a knock-on effect from queuing traffic across the boundary and can serve as a control group for northbound avenues, in which congestion north of the boundary could spill over into the control zone. We interact our previous difference-in-difference coefficient with an additional difference for the uptown vs. downtown direction of traffic along each avenue and include fixed effects at the level of treatment-zone and month, which absorb the variation that previously

identified the difference-in-difference coefficient. Southbound avenues do flow in the direction toward a denser city with slower road speed (as illustrated in the slopes of within-year spatial trends in Figs. 3.3 and 3.2), and baseline speed differences could reflect differences in a baseline level of traffic spillovers across the boundary. However, the knock-on effects from this baseline difference are controlled for by $run - bin$ fixed effects, while the addition of boro-zone interacted with time effects in the triple difference specification absorb the average differential trend in speed between north and south of the boundary that is constant across both directions of traffic. We introduce a triple difference coefficient on the interaction of the boro-zone and roll-out intensity measure and an indicator for avenues running in the uptown direction. The coefficient on this interaction is identified, as the boro program rolls-out, from the variation between traffic directions in the within-avenue speed differences across the hail-exclusion boundary: since any knock-on effect would travel backward in the direction of traffic from treatment into control area for uptown runs only, knock-on effects will cause the speed difference across the boundary to be lower for avenues running uptown relative to those running downtown.

The coefficient for the triple difference in Col. 8 is not significant (p-value = 0.145) but its magnitude is consistent with the scale of spillovers that were found in specifications 5-7 by varying the location of the control zone: we estimate a 1.1 percentage point attenuation in the uptown direction, for which road segments in the control zone are subject to spillovers, relative to the downtown direction in which they are not.[24]

Lastly, we evaluate spillovers through spatial autoregressive models in which we allow for dependence in speed between neighboring road segments. Our aim in estimating a spatial model is to decompose the impact of increased taxi activity due to the boro program into the direct effect on an affected road segment vs. an indirect effect, and in turn use

[24]If we isolate the sample to a narrower spatial window around the boundary we estimate a knock-on effect in the range of 1.6 to 2.1 percentage points (for 250 and 500 meter window around the boundary, respectively), both significant at the 1% level. We could ask whether the coefficients estimated from these specifications identify purely the queuing (or knock-on) channel for road congestion. This will depend on whether the other channels we outlined above have a balanced effect over avenues in both directions of traffic. On (ii), we see no reason why competitive retrenchment of yellow taxis or ridehail should be differentially localized to a direction of traffic. On (i), boro taxis do transport passengers southbound across the boundary, but these trips are usually matched by a "dead-head" trip back north to search for the next hail. We cannot use TLC data to quantify the fraction of southbound trips that generate dead-head returns, since boro taxi IDs are not available from the TLC, but our sample of aerial imagery shows that use was balanced in both directions of traffic (we describe this data in Section 3.5.2). Across 18 images, mostly taken outside peak commute hours, from 10 am to 2 pm, we observe 258 cases of boro taxis south of the East 96th street boundary in our area of interest in the Upper East Side of Manhattan (a density of 9.9 boro taxis per km^2), of which 184 were traveling in a north- or south-bound orientation (the remainder were parked or traveling cross-town). Of these, 96 taxis (52.2%) were bearing south and 88 (47.8%) were bearing north.

this decomposition to quantify the extent to which spillovers may bias the difference-in-difference estimation of the program's impact. Spatial models have been subject to the criticism (Pinkse and Slade (2010); Gibbons, Overman, and Patacchini (2015)) that a single reduced form (i.e. the regression of a dependent variable on lagged values of regressors) can be mapped to different sets of structural parameters depending on assumptions on the nature of spatial interactions that are hard to verify (for example, an exclusion restriction on direct effects of lagged regressors distinguishes a pure spatial autoregressive model from one with spatial lags). We do not take a stand in this debate since we are not interested in estimating spatial interaction parameters in their own right. However, there are a couple reasons why a road network is arguably an ideal application for a pure spatial autoregressive model. First, adjacency links are well defined and directed. We consider up to three possible interactions: i) cars can slow down cars that are behind them, so $s_{r,b}$ depends on $s_{r,b-1}$ if $b-1$ is ahead of b in the direction of traffic, ii) s_{rb} depends on $s_{r'b}$ if r and r' are different directions of traffic on the same avenue (note that this interaction is subject to the reflection problem if estimated in isolation) and iii) s_{rb} depends on $s_{r'b}$ if r and r' are neighboring avenues with the same direction of traffic. Second, exclusion restrictions on the structure of spatial lags are reasonable: if a car is slowed down by a double-parked vehicle or a bike lane that is 100 meters up the road, this happens by creating a bottleneck and a slowdown 100 meters ahead that is transmitted down the road through traffic speed, and not through the direct action of double-parking on a car that is 100 meters away.

Following Gibbons, Overman, and Patacchini (2015) we estimate a structural equation of the form $\mathbf{Y} = \rho\mathbf{WY} + \beta\mathbf{X} + \varepsilon$ by directly instrumenting the endogenous, spatial autoregressive term with its reduced form in terms of series of distributed lags for the exogenous variables, i.e. $\mathbf{WY} = \pi_1\mathbf{WX} + \pi_2\mathbf{W^2X} + \pi_3\mathbf{W^3X} + \pi_4\mathbf{W^4X} + \cdots + \nu$, where $\pi_i = \rho^{i-1}\beta$ and $\nu = (\mathbf{W} + \rho\mathbf{W^2} + \rho^2\mathbf{W^3} + \dots)\varepsilon$. In Table 3.7 we report model estimates using four spatial lags in the first stage and under two alternative spatial weighting matrices: Col. 1 employs a matrix $\mathbf{W_a}$ that includes all the interactions listed above, while the matrix $\mathbf{W_b}$ used for Col. 2 includes only backward effects along a direction of traffic. The estimated coefficient for spatial dependence varies between both models, which affects the precision and the

[25] If we let b indicate here an individual run and 10 meter bin, the vector of estimated total (direct plus indirect) structural effects for every run-bin is given by $\hat{\mathbf{t}} = \hat{\beta}_{DID}\hat{\mathbf{S}}\mathbf{e}$, where $\hat{\mathbf{S}} = (\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}$ and $\mathbf{e}$ is an indicator vector for the bins in the boro zone. The average direct structural effect is $\hat{\beta}_{DID}\frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}}\mathbf{s_{bb}}$ where $\mathbf{s_{bb}}$ are the diagonal elements of $\hat{\mathbf{S}}$ and $\mathcal{B}$ is the set of bins in the boro zone. The average direct plus indirect effect on the boro zone is $\mathbf{e}'\hat{\mathbf{t}}$ and the indirect effect on the boro zone is the difference between the previous two. Indirect effects on the control zone (in its entirety, or average spillovers over specified segments) are defined analogously by selecting and averaging from the vector $\hat{\mathbf{t}}$.

decomposition of program impact. Estimates of the program's total effect on the boro zone (in Col. 1) are precisely estimated and exceed the difference-in-difference estimates by about 2 percentage points.[25] Spatial dependence decays rapidly in this model and spillover effects are small, standing at about 1 percentage point 50 meters across the boundary and quickly vanishing thereafter. Estimates in Col 2 are less precise: although s.e. on the spatial dependence term are smaller in magnitude, the estimated coefficient $\hat{\rho}_b = 0.956$ is closer to unity which results in higher variance for average structural effects, as reflected in the larger bootstrapped confidence intervals for estimated total and spillover effects. The bootstrapped 95% confidence interval for the overall effect is estimated in the range of 5.6% to 6.5%, whereas spatial spillovers are estimated as a 2.68 percentage point slowdown (with a 95% confidence interval of 0.026 to 0.027) at the first 50 meters, vanishing at a slower rate to remain at 0.5 percentage points 750 meters south of the boundary (0.003 to 0.008 confidence interval). The confidence intervals for the average spillover effect from the boro zone into the hail-exclusion zone span a range from 0.1 to 1.0 percentage points across models, and we will employ these magnitudes below to evaluate the robustness of our baseline results to spatial spillovers.

We review the addition of further controls to our baseline specification in Table 3.8, including types of bike lane and "311" hotline complaints, as well as commercial zoning.[26] The bottom line from the table is that the inclusion of a large set of road level controls has a negligible effect on our estimate of the boro program's reduced-form impact: in Col 1 we present results for our baseline DID specification, but with no controls, and find $\hat{\beta}_{DID} = 0.082$. Column 2 includes a large set of road controls, Col 3 includes a commercial zoning change which came into effect in June 2012, restricting ground floor commercial frontages with the goal of "(maintaining), over time, the general multi-store character (and) promoting a varied and active retail environment" along several avenues in the Upper West Side. The addition of both jointly (presented in Col. 4) reduces the DID estimate to 0.077.[27]

The specification in Col. 2 includes a set of additional, more disaggregated controls of the alternative road uses included in our baseline (Table 3.4, Col 2), such as four types of bike lane, six categories for "311" complaints and a control both for Citibike stations

[26]New York City provides a wealth of urban data to the public, its online portal is opendata.cityofnewyork.us. See Glaeser et al. (2015) for a discussion of urban data and empirical research. A common roadblock to the use of urban data in causal research designs is that cities do not always store or maintain timestamped versions of the data that they produce, whether from administrative records or urban sensors.

[27]The commercial zoning control is a step function that changes in June 2012 over a large area in the Upper West Side. We include zoning only as a robustness check, rather than in our baseline estimates, since it has a small incidence that may be an artifact of overfitting.

Table 3.7: Reduced Form Impact of Boro Program on Congestion.
Spatial Autoregressive Models for Spillovers Across the Hail-Exclusion Boundary

	(1) Log avg. s/m ($\mathbf{W_a}$: three adjacencies)		(2) Log avg. s/m ($\mathbf{W_b}$: backward adjacency)	
ρ_a	0.671	(0.025)		
ρ_b			0.956	(0.009)
BZ x Roll-out	0.034	(0.002)	0.003	(0.001)
Bikelane	0.007	(0.001)	0.001	(0.000)
Citibike	0.000	(0.003)	−0.000	(0.001)
311 complaints	−0.000	(0.000)	−0.000	(0.000)
Pothole	0.000	(0.000)	0.000	(0.000)
Run-bin × Run-my FE	Y		Y	
Avg. direct plus indirect effect within BZ	0.1016	[0.0998,0.1039]	0.0613	[0.0555,0.0649]
Avg. direct effect within BZ	0.0366	[0.0353,0.0377]	0.0031	[0.0022,0.0038]
Avg. indirect effect within BZ	0.0650	[0.0624,0.0681]	0.0582	[0.0532,0.0611]
Avg. indirect effect (spillover) on HEZ	0.0007	[0.0006,0.0008]	0.0085	[0.0071,0.0104]
Avg. indirect effect (spillover) on HEZ, first 50m	0.0102	[0.0096,0.0110]	0.0268	[0.0260,0.0274]
Avg. indirect effect (spillover) on HEZ, first 100m	0.0008	[0.0007,0.0011]	0.0215	[0.0202,0.0229]
Avg. indirect effect (spillover) on HEZ, first 250m	0.0000	[0.0000,0.0001]	0.0143	[0.0122,0.0169]
Avg. indirect effect (spillover) on HEZ, first 500m	0.0000	[0.0000,0.0000]	0.0065	[0.0046,0.0095]
Avg. indirect effect (spillover) on HEZ, first 750m	0.0000	[0.0000,0.0000]	0.0050	[0.0033,0.0078]
Observations	395820		395820	
R^2	0.991		0.999	
R^2-within	0.72		0.98	

Note: Standard errors in parentheses on the reduced-form 2SLS coefficients of the spatial models are robust, two-way clustered at the level of run× bin and run× month, and are displayed on the right of the estimated coefficient. F-statistics from the first stage are 278.3 and 47.4, respectively. For average structural effects (direct and indirect) we display 95% confidence intervals from bootstrapping the structural parameters over 1000 simulations, re-estimating the model and the impact matrix $(\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}$.

Table 3.8: Reduced Form Impact of Boro Program on Congestion. Alternative Controls

	(1) Log avg. s/m (No controls)	(2) Log avg. s/m (Road controls)	(3) Log avg. s/m (Comm. zoning)	(4) Log avg. s/m (All)
BZ x Roll-out	0.082	0.080	0.079	0.077
	(0.004)	(0.004)	(0.004)	(0.004)
Bikelane:				
Signed route		-0.016		-0.016
		(0.009)		(0.009)
Wide parking		0.012		0.014
		(0.009)		(0.009)
Standard		0.016		0.016
		(0.006)		(0.006)
Protected		0.034		0.034
		(0.006)		(0.006)
Citibike station		-0.006		-0.006
		(0.002)		(0.002)
Citibike (south of station)		-0.004		-0.005
		(0.008)		(0.008)
311 complaint:				
Blocked driveway (no access)		0.002		0.002
		(0.001)		(0.001)
Blocked driveway (partial access)		-0.001		-0.001
		(0.002)		(0.002)
Blocked roadway (construction)		0.001		0.001
		(0.001)		(0.001)
Failed roadway repair		0.000		0.000
		(0.001)		(0.001)
Pothole complaint		-0.001		-0.001
		(0.000)		(0.000)
Rough roadway		0.001		0.001
		(0.001)		(0.001)
Pothole outstanding		-0.001		-0.001
		(0.001)		(0.001)
Zoning: EC-2/EC-3 districts			-0.020	-0.021
			(0.005)	(0.005)
Run-bin × Run-my FE	Y	Y	Y	Y
Observations	406260	406260	406260	406260
R^2	0.97	0.97	0.97	0.97
R^2-within	0.15	0.15	0.15	0.16

Note: Unless otherwise specified, robust standard errors are two-way clustered at the level of run× bin and run× month.

and the road segment to the south of these. This disaggregation has no effect on the DID estimate, and neither does a full set of interactions (not reported) for the two count variables of potholes and complaints. It is not surprising that some of these controls have no incidence on our estimates: Citibike, in particular, only expanded into our area of study after the launch of the boro program, and by June 2016 had not expanded north of W 110 St or E 96 St. The table also reports on a secondary empirical finding on the effect of bike lanes on speed: although our study is not intended to evaluate the effect of bike lanes on street speeds in New York, and our area of study does not exploit all available variation in the deployment of bike lanes, we find that the introduction of bike lanes in northern Manhattan was correlated with slower street speeds, contradicting a previous report by the NYC Department of Transportation (NYC-DOT (2014)). In our baseline specification we find that avenues become 2.4% slower following the introduction of any bike lane, whereas the breakout across bike lane types in Table 3.8, Col 2 finds that the effect is monotone in the lane type's utilization of road space: signed routes (which do not employ road space, and consist of signage indicating how cyclists should navigate an intersection) are associated to a marginally significant 1.6% speed increase. We find no effect on wide parking lanes (designed to provide additional space for cyclists), whereas standard bike lanes (which cars can invade, in a traffic violation) are associated with a 1.6% speed decrease, and protected bike lanes (in which, usually, a parking lane is moved inward to separate cars from cyclists) are associated to a 3.4% speed decrease. It might be expected that bike lanes should reduce traffic speed by reducing the maximum throughput capacity available to cars. Speed reductions may be small, however, if the flow of cars decreases due to substitution to other roads or transportation modes such as bicycles or transit.[28] Further, bikelanes may improve traffic if they isolate pre-existing cycling from other vehicles.

3.5 The Elasticity of Congestion to Taxi Supply

In the previous section we quantified the boro taxi program's impact on traffic speed in northern Manhattan. The program's effect on congestion is a first-order cost, to be evaluated by New York transportation regulators along with the program's transportation benefits such as increased taxi hail availability throughout the five boroughs.[29] In this

[28] Reduced flows following lane reductions would be consistent with the finding by Duranton and Turner (2011) that vehicle-miles traveled are proportional to highway lanes, although this finding is at the level of metropolitan statistical areas and in the long run.

[29] The TLC has issued two reports on the boro program, containing information on changes in the availability of taxi service outside the "Manhattan Core" as well as some assessment of the program's impact on traffic speed. On congestion, the first report (NYC-TLC (2013)) employs data through October 2013 and

section we estimate the change in congestion in terms of changes in taxi supply. Our goal is to provide a congestion elasticity that can be applied outside of northern Manhattan to regulation issues in the city more broadly, such as medallion deregulation[30] and the expansion in car-based transportation from "disruption" by ridehail.

TLC trip record data allows for the measurement of taxi activity at a high resolution in space and time, but only data on trip endpoints (pickups and dropoffs) is available throughout the roll-out of boro taxis.[31] Taxi pickups are the closest measure to local supply that is available in the TLC data, as they occur at the locations at which taxis cruise for street-hails. Pickups, however, are an imperfect measure of supply because they arise endogenously from matches between vacant taxis searching for hails and potential passengers attempting to hail taxis from the street curb (and who, on the margin, may choose alternative transportation modes or not to begin travel from a given location). Our interest lies in the effect on congestion from the long term increase in supply, as might result from medallion deregulation or the unregulated entry of a close taxi substitute. We contrast two approaches to estimating a curve for the congestion-supply elasticity.

3.5.1 "Back-of-the-Envelope" Approach

The elasticity of congestion to supply can be decomposed into an elasticity of congestion to pickups times an elasticity of pickups to supply:

$$\eta_{c,s} = \eta_{c,p} \times \eta_{p,s} \tag{3.6}$$

and these elasticities may vary by location or location attribute. In this section we estimate the elasticity of congestion to supply in two steps, as suggested by Eq. (3.6). First, we estimate $\eta_{c,p}$ by exploiting the roll-out of the boro program. We then rescale $\eta_{c,p}$ using

provides a table with year-on-year changes in average taxi trip speed for trips between 28 broad subdivisions of the city, using data from both yellow and boro taxis. The results are heterogeneous and inconclusive, and as of October 2013 only 1,128 boro taxis (17.3% of the full roll-out) had provided a trip. The assessment in the second report (NYC-TLC (2015)) consists of the following statement: *"As to congestion throughout all five boroughs outside of the Manhattan Core, average Boro Taxi trip speeds have remained steady over the last year, even as more Boro Taxis have entered into service. Comparing trips completed in January through June across the five boroughs, the average trip speed was 11.2 mph in 2014, compared to 11.3 mph in 2015."* This assessment is i) restricted to trips from Boro taxis only, ii) an average for most of the city, rather than between areas where the program's incidence may be expected to vary, and iii) uses January to June 2014 as a baseline for comparison, although 5,069 boro taxis were already on the road by June 2014 (3,818 as of January 2014), which is 77.5% (58.4% respectively) of the program's peak roll-out of 6,539 boro taxis in June 2015.

[30] Buchholz (2017) and Frechette, Lizzeri, and Salz (2016) evaluate deregulation counter-factuals that do not account for endogenous changes to traffic congestion and related costs.

[31] Medallion and driver identifiers, which allow a researcher to study taxi movements between trips, are only available for yellow taxis and through 2013.

a common, "back-of-the-envelope" factor obtained from a proportionality assumption between pickups in our area of study and effective supply throughout the city. We estimate both an average and a curve; the latter is identified from heterogeneous decreases in speed and increases in taxi pickups, driven by variation within the boro zone in proximity to the treatment boundary, as well as proximity to subway stations and major intersections. In Section 3.5.2 below we address the fact that pickups and supply need not be in proportion to each other.

Elasticity of Congestion to Pickups

We begin by estimating an average of the elasticity of congestion to pickups over the entire treatment zone ($\bar{\eta}_{c,p}$), i.e. we employ the boro roll-out as an instrument for the effect of pickups on travel times. The first stage of our empirical specification is:

$$\log \text{pickups}_{rbm} = \pi_{DID}\ \text{roll-out}_m \times \text{boro zone}_{rb} + \pi_x x_{rbm} + \delta_{rb} + \delta_{rm} + \nu_{rbm} \tag{3.7}$$

and the equation for the second stage is:

$$s_{rbm} = \bar{\eta}_{c,p} \log \text{pickups}_{rbm} + \gamma x_{rbm} + \delta_{rb} + \delta_{rm} + \varepsilon_{rbm} \tag{3.8}$$

where $\bar{\eta}_{c,p}$ is the congestion elasticity of pickups, and we require the exclusion restriction that, conditional on covariates, any effect from the boro taxi program's roll-out on traffic speed occurred through the activity of both yellow and green taxis (as measured by pickups) rather than any other channel.[32] The first three columns of Table 3.9 present the OLS estimate of pickups on travel time (Col 1), as well as the first and second stages (Cols. 2-3), for the entire sample. Columns 4-6 provide estimates for the east side, and Columns 7-9 the west side. OLS estimates of the elasticity are small, in the range of 0.013-0.016. First stage estimates show that the boro program's full roll-out had a substantial impact on taxi pickups north of the hail-exclusion boundary: on the order of a 76% increase for the entire sample, although 45% higher on the east side of Manhattan than the west. The instrumental variable estimate of the congestion elasticity of pickups is 10.6%, or 11.7% on the east and 8.3% on the west side of Manhattan.[33] Since our instrument is a (roll-out intensity-weighted) binary variable, the 2SLS estimator is an (intensity-weighted) Wald

[32]On avenues with two directions of traffic we use a trip's bearing (i.e. angle to the destination) relative to the avenue centerline to classify each pickup as occurring in the avenue's uptown or downtown direction.

[33]Table 3.14 presents OLS and 2SLS results using the B-spline projection travel time measures: the estimated congestion elasticity of pickups is higher in the overall results and for the east side, but not the west side. Estimates are less precise and are not significant on the west side.

estimator and we could read the coefficient for each 2SLS estimate from the ratio of the reduced-form (presented above, in Table 3.4.2 for the entire sample, and Tables 3.6.1 and 3.6.2 for the east and west sides) to the first stage estimates. Although the reduced-form impact of the program on traffic speed was double on the east than the west side, we can see that this was partly due to a differential impact on taxi activity, so the gap between congestion elasticity estimates for the east and west sides is narrower than the gap in the reduced-form impact of the program. Figure 3.13 illustrates how the boro program roll-out instrument identifies the effect of pickups on travel time.[34] Pickup levels are correlated with travel times, with both higher on average in the hail-exclusion zone. This correlation is reflected in the low magnitude of OLS estimates, which are driven by small changes over time in location-specific attributes, e.g. urban density and economic activity. The shading on the scatter plot into four cells by treatment status and period allows us to see that following the boro program's introduction there was a substantial increase in taxi activity in the boro zone, and that this occurred along with a substantial increase in travel times (in the control zone, taxi pickups barely changed, and there was a more moderate increase in travel times).

[34]The figure is a scatter plot of the log travel times over log pickups in our estimation sample, but we have averaged blocks of 10 contiguous 10 meter blocks on a run and within a given month so as to plot a tenth the number of points that are in the estimation sample. Road segments in the hail-exclusion zone are in light orange for the pre- period and dark orange for the post- period, and likewise road segments in the boro zone are in light or dark green. Axes are in log travel time and log pickups, but absolute values for speed and pickups are included for convenience.

Table 3.9: Wald Estimates of the Congestion Elasticity of Taxi Pickups

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Log avg. s/m	Log pickups	Log avg. s/m	Log avg. s/m	Log pickups	Log avg. s/m	Log avg. s/m	Log pickups	Log avg. s/m
	(OLS)	(1st stage)	(2SLS)	(OLS)	(1st stage)	(2SLS)	(OLS)	(1st stage)	(2SLS)
Log pickups	0.015		0.106	0.015		0.117	0.016		0.083
	(0.001)		(0.007)	(0.001)		(0.009)	(0.002)		(0.011)
BZ x Roll-out		0.764			0.860			0.594	
		(0.029)			(0.038)			(0.043)	
Bikelane	0.039	0.090	0.016	0.044	0.168	-0.002	0.029	-0.098	0.035
	(0.005)	(0.037)	(0.006)	(0.006)	(0.052)	(0.009)	(0.007)	(0.037)	(0.007)
Citibike	-0.035	-0.087	-0.000	-0.043	0.008	-0.003	-0.017	-0.228	0.006
	(0.008)	(0.101)	(0.013)	(0.010)	(0.135)	(0.019)	(0.004)	(0.102)	(0.011)
311 complaints	-0.000	0.004	-0.001	-0.000	0.004	-0.001	-0.000	0.006	-0.001
	(0.000)	(0.003)	(0.000)	(0.000)	(0.004)	(0.001)	(0.000)	(0.003)	(0.000)
Potholes	-0.000	0.006	-0.001	-0.003	0.024	-0.005	0.002	-0.009	0.002
	(0.001)	(0.007)	(0.001)	(0.001)	(0.011)	(0.002)	(0.001)	(0.007)	(0.001)
Run-bin and Run-m-y FE	Y	Y	Y	Y	Y	Y	Y	Y	Y
Sample	All	All	All	East side	East side	East side	West side	West side	West side
Observations	406260	406260	406260	216000	216000	216000	190260	190260	190260
R^2	0.97	0.95	0.94	0.97	0.95	0.93	0.97	0.95	0.96

Note: Unless otherwise specified, robust standard errors are two-way clustered at the level of run× bin and run× month. Weak instrument tests are not reported, but the lowest F-stat in the table (for column 9, the West side, Bspline specification) is 35,315. About 1.16% of the 10-meter avenue segments had no taxi pickups within a particular month, so we set the dependent variable at log(pickups + 0.1). Shifting the dependent variable by 0.01 instead changes the main coefficient in Col (2) from 0.1040 to 0.1012.

We are interested in estimating heterogeneous effects in terms of location attributes (such as an initial level of supply) so that we can derive a congestion elasticity curve and apply it outside of northern Manhattan. Although it is feasible to condition on alternative location attributes (e.g. population density, proximity to transit) we estimate an elasticity coefficient for each 10 meter bin north of the hail exclusion boundary, and we will describe the heterogeneity in these elasticity estimates below in terms of the baseline levels of taxi pickups. We estimate a bin-level congestion-pickup elasticity $\eta_{c,p}(b)$ by estimating Eqs. 3.4 and 3.7 with an interaction for 10 meter bin north of the hail-exclusion boundary, and then computing Wald estimates as a ratio of the coefficients for each bin. This procedure yields a distribution of heterogeneous estimates $\hat{\eta}_{c,p}(b)$ for the elasticity of congestion to pickups.[35]

Elasticity of Pickups to Supply: Back-of-the-Envelope

The simplest approach to computing supply elasticities from pickup elasticities is to assume that pickups and supply vary in proportion to each other. As a benchmark, we rescale the estimates of $\hat{\eta}_{c,p}(b)$ by constant values for $\eta_{p,s}$ obtained by assuming proportionality between levels of pickups and an effective taxi supply. In particular, we assume that the density of vehicles of type $v = \{yellow, boro\}$ in zone $z = \{hail\text{-}exclusion, boro\}$ and in period $t = \{Before\ Aug.\ 2013, After\ Aug.\ 2013\}$ is given by $(1/Area\ in\ km^2_z) \times (Avg.\ medallions\ per\ day_{vt}) \times (Avg.\ hours\ per\ medallion\ per\ day_{vt}/18) \times (Pickups_{vzt}/Total\ pickups_{vt})$. Vehicle densities obtained from this back-of-the-envelope formula are displayed in Table 3.10, Panel A.[36] We then rescale $\hat{\eta}_{c,p}(b)$ by $\bar{\eta}^{BoE}_{p,s}$ (i.e., the ratio of the average change in our "back-of-the-envelope" measure taxi supply for both yellow and boro taxis over the average growth in pickups) to obtain $\hat{\eta}^{BoE}_{c,s}(b)$. By construction, all heterogeneous pickup elasticities are rescaled by the same factor. Figure 3.6 plots the set of impact estimates in blue, against baseline levels of taxi pickups prior to August 2013, along with a local polynomial regression and a bootstrapped 95% confidence interval. We discuss this curve below.

3.5.2 Measuring Changes in the Spatial Distribution of Taxi Supply with Aerial Orthoimagery

In this section we develop an approach to measuring supply (and, implicitly, the elasticity $\eta_{p,s}$) that does not require us to assume proportionality between pickups and

[35] We restrict this analysis to the east side of Manhattan due to data acquisition and processing costs for the sample of aerial imagery that we will describe below.

[36] Averages of monthly statistics are obtained from the TLC's aggregated monthly reports at nyc.gov/html/tlc/html/technology/aggregated_data.shtml, since vehicle identifiers that would allow for

Table 3.10: Vehicle Densities (per km^2). Selected Areas of Interest.

Panel A: Back-of-Envelope proportional allocation	**Yellow taxi**	**Boro taxi**	**Taxi Total**
Hail-exclusion zone			
Before Aug 2013	237.3		237.3
Since Aug 2013	232.9	1.0	233.9
	-1.8%		-1.4%
Boro zone			
Before Aug 2013	36.1		36.1
Since Aug 2013	32.5	36.1	68.6
	-10.0%		90.1%

Panel B: Aerial counts	**Cars**	**Black cars**	**Trucks**	**Yellow taxi**	**Boro taxi**	**Total**	$\widehat{\textbf{Private}}$	$\widehat{\textbf{FHV}}$
Hail-exclusion zone								
Before Aug 2013	264.4	108.7	82.0	247.0		684.1	269.8	85.3
Since Aug 2013	270.4	164.9	67.1	205.4	10.1	717.8	289.6	145.7
	9.7%	51.7%***	-18.2%	-16.8%***		4.9%	7.3%	70.8%**
Boro zone								
Before Aug 2013	286.9	126.0	65.1	61.8		539.8	314.5	98.4
Since Aug 2013	309.8	160.0	54.0	58.2	36.0	617.9	346.6	123.2
	8.0%	27.0%*	-17.1%	-5.8%		14.5%	10.2%	25.2%
Midtown								
Before Aug 2013	370.3	153.5	86.1	530.7		1140.7	411.4	112.5
Since Aug 2013	404.9	341.8	79.3	468.2	7.6	1301.8	383.5	363.3
	9.3%	122.7%***	-7.9%	-11.8%*		14.1%	-6.8%	222.9%***

Note: "Hail-exclusion" and "boro" zones in both panels designate the area of interest in northeastern Manhattan, all areas described in 3.8.2. Rows with percentage changes in Panel B are labeled with stars (* $p < 0.1$, ** $p < .05$, *** $p < 0.01$) for a two-sided difference in means test for the vehicle density (in levels) between the period before August 2013 (11 scenes) and after (15 scenes). Almost all scenes were captured around noon due to favorable lighting conditions, and scenes include both weekday and weekend dates (see 3.8.2 for details). For each zone, period and vehicle type we estimate a weekend factor and re-weight weekend densities accordingly. Growth rates between the averages of vehicle densities that are weighted or unweighted in this manner are comparable, but unweighted densities are on average 14.3% lower due to lower counts in the weekend sample.

supply. We seek to avoid this assumption because the ratio of pickups to supply may well change under a large and localized supply shock, i.e. the otherwise unobserved event that allows us to identify the effect of taxis on traffic speed.

Our approach is to directly observe changes in supply by constructing a new dataset of taxi and other vehicle locations over space and time. We acquire aerial orthoimagery for northeastern Manhattan captured over 29 dates, 11 before and 18 after August 2013. We manually digitize, i.e. tag, the observed location of 131,921 vehicles in this imagery.[37] Figure 3.5 shows the densities of taxi locations on the north-east side of Manhattan before and after the boro program.[38]

Aerial imagery reveals how the supply of taxis reallocated around the hail-exclusion boundary following the introduction of the boro program: yellow taxi density decreased slightly throughout most of the kilometer south of the hail-exclusion boundary. At the boundary itself there has been a spike in supply from boro taxis, but this is almost exactly offset by a retrenchment from yellow taxis. The density of yellow taxis has remained largely unchanged north of the boundary, while the additional supply of boro taxis has led to increases in the taxi density from both types of taxi to average 64.6% over the two kilometers north of the boundary (and ranging from 0.0 to 291.5% growth). We calculate a distribution of heterogeneous elasticity estimates $\eta_{c,s}^{Aerial}(b)$ using the bin-level reduced form changes in speed in the numerator and changes in vehicle densities in the denominator. Since our estimates of changes in speed come from a difference-in-difference specification, we analogously define the denominator of the congestion elasticity term for bin b as $(log\ d_{t',b} - log\ d_{t,b}) - (log\ d_{t',hez} - log\ d_{t,hez})$ where t and t' indicate the average

the computation of daily or hourly vehicle utilization are not available for boro taxis or for yellow taxis after 2013. The value of 18 hours in the denominator for vehicle utilization reflects the fact that the modal yellow taxi medallion was employed in two shifts of 9 hours each in 2013. Boro taxis are used less hours in a day then yellow taxis, so adjustment by an hourly utilization rate results in a lower effective supply of boro relative to yellow taxis, given the same number of outstanding medallions. The choice of 18 hours is somewhat arbitrary and affects the levels of vehicle densities, but not rates of change or elasticity factors.

[37] In our analysis, however, we only employ 26 scenes through June 2016, because the TLC has anonymized geographic coordinates since the second half of 2016. Sources and methods are detailed in 3.8.2. We employ all the aerial orthoimagery available for New York City from all known sources (public or commercial) but one, the exception being a second commercial vendor. Examples of imagery and tagging are displayed in Figures 3.7a and 3.7b. We evaluated and disregarded satellite imagery of the highest resolution available (Digital Globe WV2 and WV3, Airbus Pleiades, Kompsat-3), which is unsuitable for vehicle roadbed counts due to resolution and visual obstruction from buildings. See Donaldson and Storeygard (2016) for a discussion of satellite imagery in economics research.

[38] We have overloaded the term "density" here, in the sense that the plotted curves are kernel density estimates of observed vehicle locations along the centerlines of Manhattan's north-south avenues, but we rescale the density estimates so that the y-axis indicates the density (in vehicles per km^2) that any segment length over the x-axis would mean-integrate to over the width of our area of interest, which is between 1st and 5th Avenues. Rescaling allows us to compare changes in vehicle density along a north-south axis before and after the introduction of the boro program.

of densities over images before and after the boro program's introduction, and the second difference is in the logs of the densities provided in the hail-exclusion zone rows of Table 3.10, Panel B. Note that we compute the denominator of congestion elasticity estimates from changes in taxi counts, and not changes in counts for total vehicles on the road. Estimates of bin-specific speed changes are a reduced form of the net impact from both an increased taxi supply and any countervailing substitution, while any local substitution behavior toward private vehicles would partially offset the change that we measure in taxi counts, leading us to overstate the net change in vehicles on the road and understate the congestion elasticity from net changes in vehicles on the road. Our congestion elasticity estimates should therefore be interpreted as a reduced form that includes the adaptation in private driving behavior that took place in the setting of boro taxi deregulation, and therefore understates the structural congestion elasticity of taxi supply that would obtain under a counterfactual in which private car use remains constant.[39] In Section 3.6 below we describe a methodology to attribute non-taxi car counts to private vs. for-hire-vehicles. Applying this methodology to compute the last two columns of Table 3.10.B we do not find increased substitution of taxis for private cars in the boro zone, but we do find that the boro zone has seen relatively less growth in for-hire-vehicles.

In Figure 3.6 we plot in orange the set of heterogeneous congestion elasticities against baseline levels of taxi pickups prior to the boro program (along with a local polynomial regression and a bootstrapped 95% confidence interval). The gap between the congestion elasticity curve estimated from aerial imagery and the back-of-the-envelope curve arises precisely due to the variation in the pickup-to-supply elasticity that aerial imagery is intended to control for: the ratio of the growth in pickups relative to the growth in supply has been larger in areas that were previously less well served with taxi supply, i.e. had lower baseline taxi density. The narrowing gap at higher levels of density can be explained if – as we would expect – the match efficiency between taxis and potential passengers increases with density. Accordingly, the back-of-the-envelope estimate derived from the number of pickups (i.e. matches) does a better job of approximating changes in supply at higher densities.

Employing our preferred aerial-imagery measures of supply, we find that the congestion elasticity of taxi supply averages 0.2 and is at 0.3 at the 90th percentile of the curve,

[39] We are unable to estimate a congestion curve for total car counts because some intervals slowed down yet show infinitesimal changes in total car counts, leading to a high variance in elasticity estimates: 10% of the area north of the boundary saw a change of less than 1% in total car counts, whereas the smallest bin-level change in taxi supply is a 12.8% increase. The ratio of the reduced-form impact estimate (an 8.0 percentage point slow down) to the differences in log total vehicle densities (an 8.7 percentage point difference-in-difference growth in total vehicle density, most of which is driven by growth in boro taxis) yields an average travel time to vehicle density elasticity of 0.91.

as the curve increases with density and is somewhat convex. We highlight our congestion elasticity estimates at the 90th percentile of baseline taxi pickups because areas with higher taxi activity in our sample are the most comparable to average levels of taxi activity in midtown Manhattan.

3.6 Application: the Role of Ridehail in the Slowdown of NYC Traffic

In this section we apply our estimates of the congestion elasticity of taxi supply to explain the substantial slowdown of New York City since 2013. First, we provide descriptive evidence on the depth, timing and extent of the traffic slowdown throughout the city's streets and highways since 2013. Next, we discuss the two existing studies that evaluate the effect of ridehail on traffic congestion in the city, and we complement our assessment of these studies with evidence on traffic speed and vehicle use. We then answer a simple accounting question: given our estimates for the congestion elasticity of taxi supply, could the city's slowdown be accounted for by the additional supply of for-hire vehicles that was enabled by ridehail applications (e.g. Uber, Lyft, Via, Juno and Gett)? To answer this question, we employ aerial imagery counts to estimate the growth in the supply of for-hire vehicles within an area of midtown Manhattan, as well as measure substitution from private vehicles. We find that most of the observed slowdown in midtown Manhattan can be explained by growth in the supply of for-hire vehicles.[40]

Figure 3.1 shows that the travel time of the median yellow taxi trip has increased substantially since 2013: after remaining flat at an average 5.1 minutes per mile from June 2009 to June 2013, from 2013 onward the median travel time started to rise and has exceeded 6 minutes per mile since the second half of 2016.[41] Median or average travel time over millions of monthly trips could be subject to selection from changing patterns in the use of taxis: congestion could possibly remain constant while passengers choose to take

[40]To our knowledge, our study is the first to quantify growth in the supply of vehicles on the road due to ridehail, and to provided direct (i.e. non-survey) evidence on substitution from private vehicles. Uber has provided academics with data for the purposes of studying the value of the platform to consumers and workers, e.g. Cramer and Krueger (2016), Cohen et al. (2016), Hall and Krueger (2016), Angrist, Caldwell, and Hall (2017), Chen et al. (2017), Castillo, Knoepfle, and Weyl (2017) and Hall, Horton, and Knoepfle (2017). In particular, Castillo, Knoepfle, and Weyl (2017) illustrate how dynamic pricing can allow for increased density of Uber vehicles, increasing Uber's efficiency by reducing travel to (and time to) pickup. The paper does not address the impact of Uber on road use or speed. To date, Uber has not provided access to data for the study of congestion externalities or the degree to which ridehail complements or substitutes transit or the use of private vehicles. Outside of a data agreement, Ge et al. (2016) find evidence of racial discrimination on the platform.

[41]The same pattern shows up in survey responses from the US Census' American Community Survey (ACS). Manhattan residents reported that travel time to work by car remained approximately constant between 2009 and 2013, but increased by 2.75 minutes (7.6%) from 2013 to 2016.

taxis on more congested roads, driving up the median travel time.

To condition on possible changes in travel patterns over space, the plots in Figure 3.8.(a)-(c) show time series for selected percentiles of taxi travel time per mile during the month of June, with each line tracking a percentile for a particular "taxi-zone" pair within midtown Manhattan.[42] Percentiles of the travel time across most origin-destination pairs remained flat until June 2013 and have increased thereafter. Figure 3.8.(d) summarizes the variation between June 2013 and June 2016: across 342 origin-destination cell-pairs within midtown Manhattan every percentile has shifted upward on average; the median and 10th percentile of travel time increased in 98.9% of all pairs, whereas the 90th percentile increased in 95.3% of pairs. The median travel time has increased by 15.2% on average across all origin-destination cell pairs in midtown. We describe congestion in midtown for the purposes of our accounting exercise but while midtown slowed down by more than average, it is not unusual: extending the previous comparisons to all taxi-zone pairs with more than 5 trips in both June 2013 and June 2016, we find that the median taxi trip slowed down by more than it did on average in midtown across 34.4% of the cell-pairs in the city. Figure 3.9 reports on speed changes using a data source that is novel to our paper: pair-linked event data from a network of EZ-pass sensors throughout highways and expressways in the city. There is a gap in the data around 2014, but there has been an evident slowdown since 2013 throughout the entire distribution of highway speeds (fig. a), throughout the day (b), and throughout the city's network of highways (contrasting (c) and (d)).[43]

The fact that the speed of traffic in New York City has declined substantially in recent years has been reported previously in the media and through aggregate statistics in city transportation reports.[44] However, there is no prior academic research on the timing and scale of the slowdown, or credible empirical evidence on its causes. In January 2016 the Office of the Mayor issued the ad-hoc "*For-Hire-Vehicle Transportation Study*" (NYC (2016)), which concluded that "*reductions in vehicular speeds are driven primarily by increased freight movement, construction activity, and population growth. (...) E-dispatch (...) did not drive*

[42] "Taxi-zones" are polygonal cells defined by the Taxi and Limousine Commission, and are the finest spatial unit at which trip origination data is provided by ridehail apps. From July 2016 onward the TLC has also started anonymizing individual taxi trip records to the taxi-zone level. There are 19 taxi-zones in midtown Manhattan and 263 in all of the city.

[43] While the NYC Department of Transportation exposes the data to the internet as a real-time feed (see Data 3.8.1), the DOT has claimed in correspondence with us not to have archived it. We have archived the data since 2015, and data from three months including June 2013 was kindly made available by Prof. Tomonari Masada of Nagasaki University, who archived it for unrelated research.

[44] Examples include Fitzsimmons and Hu (2017) and Aaron (2015). NYC-DOT (2016) reports a 12% slowdown between 2010 and 2015 in the annual average of taxi trip speed in Manhattan below 60th Street, and a 20% slowdown in a "Midtown Core" area.

the recent increase in congestion in the CBD." The report fails to meet modern standards for identification and there is evidence that it was produced under strong industry lobbying.[45] A second report (Schaller (2017)) cautions that vehicle miles traveled in the city are on the rise due to ridehail.[46] As corroborating evidence, we estimate time series for VMTs per day, making explicit methodological decisions to address censoring in the data but remaining agnostic on the distribution of VMTs between traditional car services and ridehail (see 3.8.5). We plot time series for estimated VMTs in Figures 3.10d and 3.10e. We find that the sum of taxi (yellow plus boro) and FHV miles increased by 886 million miles (34.6%) between 2013 and 2016. Our results are broadly consistent with Schaller (2017), although that report may understate growth in miles through 2016 as well as ridehail's impact through 2015.[47]

In addition to documenting the extent and timing of speed changes and vehicle use through data from taxi meters, EZ-pass and odometer inspections, our contribution to the study of the causes of New York City's slowdown consists of i) measuring the growth in the supply of for-hire vehicles on the road, and ii) applying our estimate of the marginal congestion of taxi supply to evaluate whether increases in vehicle supply are consistent

[45]During the summer of 2015, NYC Mayor Bill de Blasio suggested capping growth in for-hire vehicles. In the words of former Uber lobbyist Bradley Tusk, "*we were able to, through a really aggressive campaign, beat back completely (...) some limitations on Uber's growth. (...) We ran $4 million in TV spots. We did radio ads. We did direct mail. We had digital ads. We mobilized our customers (...) We had five different lobbying firms. We worked every single editorial board in the city. We worked the columnists and the pundits. It was just an all-in, 24-hour-a-day campaign*" (quoted from Mehta (2016)). The claim that ridehail did not cause congestion is made by incorrectly equating trip counts to vehicle supply: "*increases in e-dispatch trips are largely substituting for yellow taxi trips in the CBD. Because these e-dispatch trips are substitutions and not new trips, they are not increasing VMT (Vehicle Miles Traveled)*." Congestion is attributed to construction and pedestrians by appealing to correlation with city-wide growth rates in these activities since 2009, a baseline year affected by economic recession.

[46]Using odometer inspection data from for-hire vehicles and taxis, and under some explicit assumptions on private car use and substitution, Schaller (2017) estimates that ridehail led to an additional 600 million vehicle-miles traveled in 2016 relative to 2013, a 27.1% growth over yellow taxi and FHV miles in 2013. Schaller's analysis requires unstated assumptions to handle the fact that vehicle mileages are censored at the last available inspection date, and to allocate flows of miles at the VIN level to stocks of ridehail vs. traditional car services.

[47]Following an initial freedom-of-information request we found that the odometer data that was provided to Schaller contained gaps. Consultation with the TLC allowed us to obtain additional inspection data. More substantively, Schaller deducts the decline in yellow taxi VMTs from his gross estimate of ridehail VMTs, under the assumption that all losses in taxi miles were substitution and not additional miles created by ridehail. However, the report does not track VMTs by boro taxis, which as Figure 3.10d shows more than offset the declines in yellow taxi VMTs though the first half of 2016. Since Schaller (2017) implicitly assumes no substitution from yellow to boro taxis, any degree of actual substitution would imply that the estimate of ridehail's growth is biased downward. The report also claims to be consistent with the finding in NYC (2016) that ridehail was not a major source of congestion through the first half of 2015. Unlike the rest of the report, however, this claim is based on growth in trip counts rather than VMTs. Analyzing VMTs for the same period as NYC (2016), we find that VMTs from yellow taxis plus FHVs increased 8.4% between June 2013 and June 2015, and by a further 17.8% between June 2015 and June 2016 (15.0% and 24.0%, respectively, if VMTs from boro taxis are included).

with the observed slowdown.

There are challenges to the use of aerial imagery to count supply on the road. Due to the cost of digitizing the imagery we restrict our analysis to an area on the east side of midtown Manhattan, delimited by 1st and 6th avenues and by 38th and 57th streets. We must also rely on a small sample of historical images captured before and after the introduction of ridehail. As in Section 3.5.2, we compare periods before and after August 2013: comparisons between these two periods are differences in means over a sample of 26 images, 11 before and 15 after August 2013 (details in 3.8.2). A further methodological challenge is that aerial images do not reveal whether observed cars are private or for-hire vehicles. We address this shortfall in the data by counting (in addition to taxis and trucks) the cars that have a black exterior color separately from cars of any color other than black. Counting by color was motivated by a prior that black is the most popular color among for-hire vehicles in New York City, as well as a growth in the share of black cars over successive aerial images that is evident to the naked eye. We then obtained current and historical snapshots of the vehicle registration data for every vehicle in the state from the New York Department of Motor Vehicles. The data contains exterior paint color and county of registration for almost all vehicle identification numbers (VINs). We calculate the share of vehicles with a black exterior paint color among those with an active NY state registration and registered as for-hire vehicles with the Taxi and Limousine Commission ($\alpha_t^{b,F}$). Equivalently, we calculate $\alpha_t^{b,P}$ for private cars (i.e. non-TLC registered) in NYC. Black-paint shares averaged 20.9% for private cars and 73.0% for FHVs in our period.[48] For each image we obtain a pair (c_i^b, c_i^{nb}) of car counts by color. If we are willing to assume that the paint color shares of cars counted in our midtown area of interest are proportional to color shares in the vehicle registration data for New York City, then we can derive an estimate of the otherwise unobserved counts for private and for-hire vehicles $(\widehat{c}_i^P, \widehat{c}_i^F)$ in midtown by inverting

$$\begin{bmatrix} c_i^b \\ c_i^{nb} \end{bmatrix} = \begin{bmatrix} \alpha_t^{b,F} & \alpha_t^{b,P} \\ 1-\alpha_t^{b,F} & 1-\alpha_t^{b,P} \end{bmatrix} \times \begin{bmatrix} \widehat{c}_i^F \\ \widehat{c}_i^P \end{bmatrix} \tag{3.9}$$

for each image.

Average vehicle densities per km^2 were reported earlier in Table 3.10, Panel B, but we refer now to counts for midtown Manhattan. We find that the number of vehicles per

[48]Vehicle counts likely include commuters from outside the city. We cannot trace the origin of the vehicles in our midtown counts to neighboring states and counties, and lack vehicle registrations by color for neighboring states. However, black-paint shares for private vehicles barely change (average 19.2%) if we include registrations from neighboring Nassau and Westchester counties in NY state.

km^2 increased by 14.1% from the 11 images prior to the 15 images after August 2013. We underscore that due to the small size of our sample and the variance over images, the estimates are imprecise: although 14.1% growth is an economically significant magnitude, it is not statistically significant. We do find a statistically significant 11.8% decline in the number of yellow taxis. We also find a 42.5% increase in the number of cars, which breaks down into a 122.7% increase in black cars and a 9.3% increase in non-black cars. Using the share of car colors in the general population of NYC vehicles, we estimate that private cars on the road decreased by 6.8%, whereas the density of FHVs increased by 222.9%. The increase in FHVs from 112.5 to 363.3 per km^2 more than offset the drop in 62.5 observed taxis per km^2: the sum of taxis plus FHVs on the road in midtown has increased 30.5% in the average of aerial images observed over the past three years, relative to the four years earlier.

In Section 3.5 we non-parametrically estimated a congestion elasticity curve for taxi supply that increased in baseline taxi activity. A purpose for estimating a curve rather than an average was to extrapolate from the northeast of Manhattan to the denser midtown area: we select an estimated elasticity of 0.304, which corresponds to the 90th percentile of baseline taxi activity. At this value, the observed 30.5% increase in taxi plus FHV density can account for a 9.3% slowdown in traffic in midtown Manhattan. Taking the median change in speed across midtown taxi zones from June 2013 to June 2016 (a 15.2% median slowdown) as a representative value, we find that the growth in the supply of vehicles on the road that was caused by ridehail can explain 61.8% of the traffic slowdown in midtown.[49]

Our application does not establish a causal effect from ridehail on traffic speed. However, it is reasonable to find (given our causal estimates of the congestion elasticity of supply) that the largest change to road transportation in New York City over the past three years can account for most of the observed slowdown. Our count (and other) data also shed light on some of the alternative causes that have been suggested for the slowdown. We cannot measure the use of the road by couriers and delivery services but, as elsewhere in the city, we find that truck counts have decreased. We count 7.6 boro taxis per km^2 after August 2013, which relative to a total count of 1301.8 vehicles suggests that boro taxis' impact on midtown congestion has been negligible. We do not adjust our estimate of the change in vehicle supply for changes in private car use, since adaptation

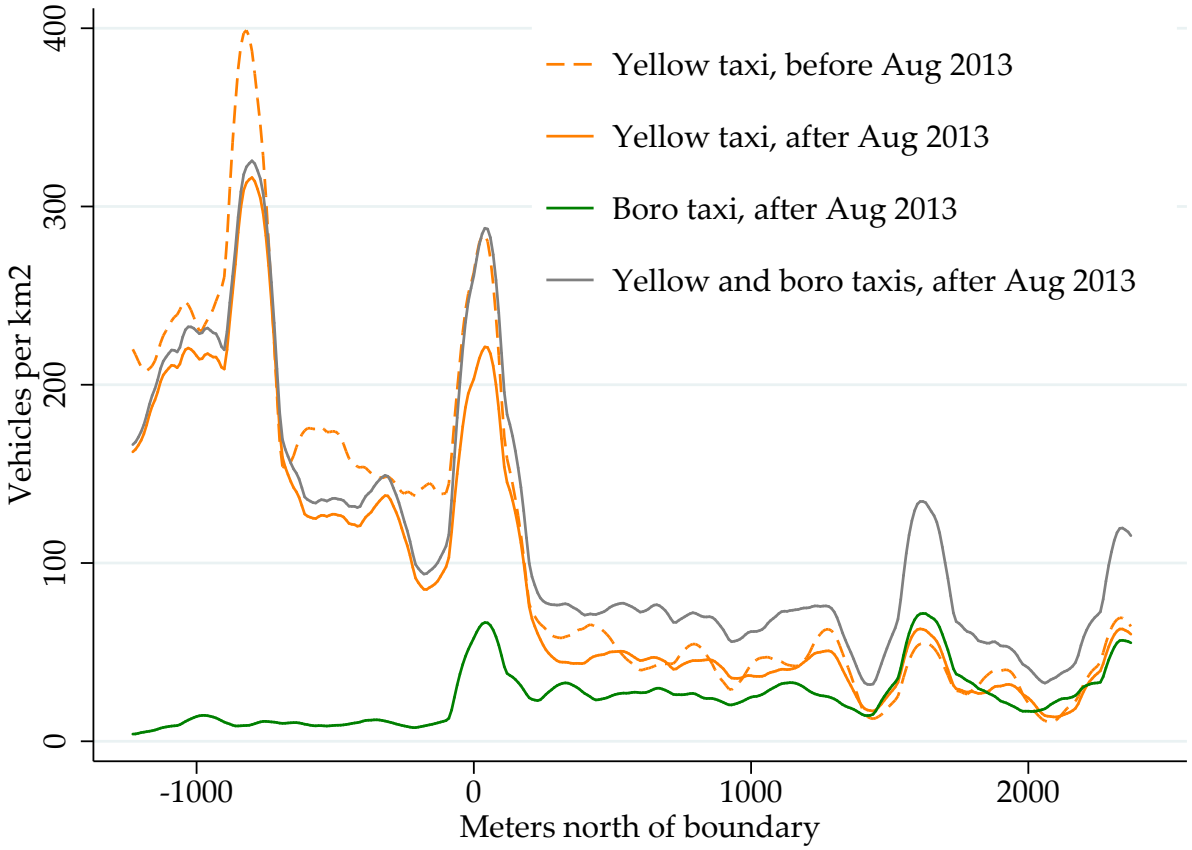
[49]We have made some conservative data decisions to arrive at this headline result: our estimates employ average speed measurements as the outcome variable (vs. B-spline projections) and do not adjust for spatial spillovers, both of which would result in higher congestion elasticity estimates. Congestion impact estimates using speed measurement only at the times that correspond to the capture times for the aerial imagery do not significantly affect results (see Table 3.15).

in private use is part of the reduced form impact in our congestion elasticity estimates. However, our aerial counts find evidence of modest substitution. Our estimates of this substitution are not statistically significant, although we emphasize that this is due to our small sample, rather than an insignificant economic magnitude.[50] Bike lanes are also unlikely to play any role in the traffic slowdown within the area of our application in midtown: the area has 46.1 km of roads and 6.9 km of bike lanes, but almost all these bike lanes were already in place by June 2013. The only bike lane expansion over the next three years was the conversion of 489 meters along 1st Avenue from shared to protected bike lanes. This expansion implies a 1.1% change in road exposure (absent a correction for spatial spillovers) to a differential slowdown of which our best estimate is 1.6 percentage points, as reported in Table 3.8. The Citibike bikeshare system was also already in place in midtown by June 2013.

It is also worth noting that the slowdown in highways and expressways that we document using EZ-pass data is inconsistent with some of the suggested alternative explanations for congestion. Bike lanes, cycling and pedestrians are almost surely unrelated to slower highways. Although courier and delivery services may have led to increased highway use, their impact would be relative to baseline utilization for delivery to brick-and-mortar stores. The channel that NYC (2016) suggests for the impact of delivery services, double-parking, is not relevant on highways either. We have not undertaken a study of highway construction projects in New York City, but the widespread spatial distribution of the highway slowdown seen from Figure 3.9c to 3.9d suggests that individual construction projects are an unlikely source. On the other hand, we cannot rule out that increases in traffic due to boro taxis, which are a negligible cause of congestion in midtown Manhattan, may have contributed to highway congestion in the outer boroughs.

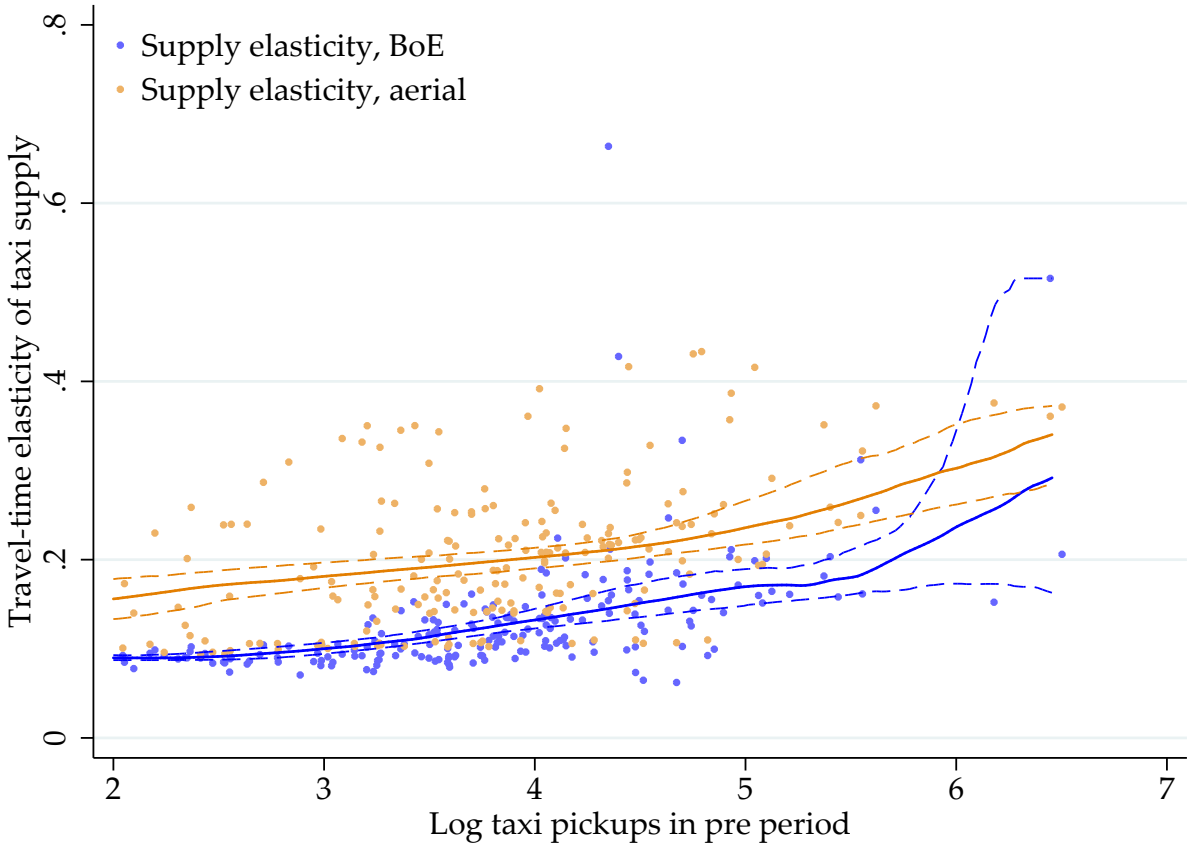
[50]There is almost no prior research on substitution patterns between ridehail and other modes. Our results are consistent with an internet survey of ridehail users in seven cities conducted by Clewlow and Mishra (2017), who report that "49% to 61% of ride-hailing trips would have not been made at all, or by walking, biking, or transit."

Figure 3.5: Densities of Taxis Located from Overhead Aerial Imagery



Note: We plot the densities of the distance between taxis on the road, as located over the area of interest in aerial imagery scenes, and the hail-exclusion boundary. Our area of interest is 4.4 square kilometers (1087 acres) on the east side of northern Manhattan, delimited by 1st Ave and 5th Ave and between East 82nd Street and East 125th Street. We count an average 522.7 yellow taxis per scene (or 118.9 per km^2) over 11 scenes prior to the boro taxi program, and plot the density of the locations of these taxis with a dashed yellow line. Following the launch of the boro program we count an average 440.5 yellow taxis and 112.1 boro taxis per scene over 15 scenes through June 2016 (125.6 taxis per km^2). We plot the density of the locations of these taxis with solid yellow and green lines. All kernel density estimates are rescaled so that they would integrate to the average number of vehicles per km^2 over 1000 meters on the horizontal axis. In this manner, the vertical axis represents a local vehicle "density" (i.e., in terms of vehicles per km^2). We plot the sum of the densities for both types of taxis in the post period with a solid grey line to facilitate before-after comparison. All densities are estimated with upper and lower bounds computed by renormalization, and congestion elasticity estimates employ only the range from 1km south to 2 km north of the boundary. Details on aerial imagery are in 3.8.2.

Figure 3.6: Congestion Elasticity Curves



Note: Elasticity estimates and curve are in blue for the "back-of-the-envelope" method and in orange for aerial count data. Points represent individual elasticity estimates per 10 meter bin in the east Manhattan boro-zone, placed on the vertical axis, and are sorted on the horizontal axis by the log of average taxi pickups from January 2009 to August 2013. Solid lines are local polynomials for the conditional mean of elasticity on average pickups. Confidence intervals are estimated pointwise, as the 97.5% and 2.5% percentile envelopes from 1000 bootstrapped samples of the local polynomial conditional mean function. Bandwidth for estimation and for every bootstrap sample is selected automatically by a default rule-of-thumb method for the asymptotically optimal constant bandwidth.

3.7 Conclusions

In this paper we evaluate the congestion cost of a particular taxi deregulation episode in New York City. We also quantify a long-run traffic congestion externality that was previously unmeasured and yet is of first order importance to the regulation of car-based transportation providers in the city. Additionally, we document that there has been a substantial slowdown in NYC's traffic since 2013, and find that the increase in supply due to ridehail can account for most of this slowdown. This is an important result both for New York City and for other cities throughout the world that are assessing the costs and benefits of ridehail and evaluating strategies for its effective regulation.

To execute this study we collect and analyze a large number of sources of urban data, of which some are "big", some have not been collected or analyzed previously, and some are entirely new. The City of New York is a leading example in the practice of making urban data publicly available. However, these efforts are at an early stage: standards have not fully developed, and are often geared toward mobile application developers rather than researchers. Some data remains siloed, and some valuable data is lost when it is deemed ephemeral by internal government users and is not archived. Additionally, changes in technology have led to private firms generating and owning data on their utilization of public goods and congestible resources. Regulators may want to develop frameworks so that businesses that are built on the use of public infrastructure report data on this use. The availability of urban data will increase in the near future, along with the use of internet-connected devices and other urban sensors. The use of methods from empirical economics in the analysis of such data can be expected to increase, motivated by changes to the economic activities that take place within cities (in transportation, these may include ridehail, bike lanes, bikeshare, dynamic tolling and autonomous vehicles). In the future, in-vehicle GPS may allow for first-best remedies such as road-specific congestion pricing – as long as regulators can quantify and price the relevant externalities.

3.8 Appendix

3.8.1 Data Sources

Taxi trips. Taxi trip records are from the New York City Taxi and Limousine Commission's Taxicab Passenger Enhancements Project (TPEP) and Street Hail Livery (SHL) Passenger Enhancements Project (LPEP). The paper uses trip records from 2009 through the first half of 2016, obtained from multiple FOIL requests by the authors and as of September 2015 available for direct download from the NYC TLC's website at: nyc.gov/html/tlc/html/about/trip_record_data.shtml. As of the second half of 2016 the TLC has started removing trip endoint coordinates from their data releases, which precludes the street speed reconstruction methods developed in this paper. Data from ridehail and other for-hire vehicle providers was also obtained by FOIL prior to September 2015, and records with the time and "taxi-zone" of pickup are available from 2015 onwards at the TLC website.

311 complaints. Data from reports to New York City's 311 complaint hotline are available at www1.nyc.gov/311/our-data.page. Each record contains the date, location coordinates, and a category classification for the complaint. This paper utilizes all 311 records in northern Manhattan for categories that may be related to the flow of traffic. We count the number of complaints per month and segment of road that fall under the following categories: "Blocked Driveway - No Access," "Blocked Driveway - Partial Access," Street Condition - Blocked - Construction," "Street Condition - Failed Street Repair," "Street Condition - Pothole," and "Street Condition - Rough, Pitted or Cracked Roads."

Bike lane maps. The New York City Department of Transportation provides a shapefile for all bike lanes, including the date of installation or modification, at www.nyc.gov/html/dot/html/ about/datafeeds.shtml#bikes. All changes in bike lanes (either new construction or modification) from 2009 through the first half of 2016 are matched to street segments. The various types of bike lane changes in the data include bike-friendly parking, protected paths, signed routes, and standard bike lanes.

Citibike system. Bike trip data is from the Citibike system data page at citibikenyc.com/system-data. Data on location and availability of bike stations was archived from the Citibike live station feed, available at citibikenyc.com/stations/json as well as the system data web page.

Potholes. Street Pothole Work order data is available at www.nyc.gov/html/dot/html/about/datafeeds.shtml#construction. This dataset contains the reported date and closed

date of pothole work orders, identified at the level of street sections between intersections, from 2010 through 2016. Each order may refer to more than one pothole.

EZ-pass traffic speed. Data for traffic sensors in New York City is made available as a real-time feed by the NYC Department of Transportation. A link to the real time feed is available at nyc.gov/html/dot/html/about/datafeeds.shtml#realtime. The NYC-DOT has claimed in private correspondence not to archive the output of this feed. From May 27, 2013 to August 20, 2013 the feed was continuously archived and kindly made available by Prof. Tomonari Masada of Nagasaki University. Data from April 2015 through August 2015 (and beyond) was continuously archived by *βetaNYC*, a civic technology non-profit, and is available at data.beta.nyc/dataset/nyc-real-time-traffic-speed-data-feed-archived. Data from August 2015 onward was also continuously archived by the authors. Additional snapshots of the data were obtained from archive.org at 11 arbitrary times and dates in 2014 and prior to 2013.

New York State vehicle registrations. Obtained from current and previous archived snapshots of the NY State Department of Motor Vehicles database for Vehicle, Snowmobile and Boat registrations, available at data.ny.gov. Data contains county, zipcode, car brand and body type, model year, exterior color and VIN for every vehicle registered in the state of New York.

Taxi and For-hire-vehicle inspection data. Obtained by FOIL from the Taxi and Limousine Commission's Vehicle Inspection Program. Data contains odometer readings and inspection dates for yellow taxis (performed quarterly), boro taxis (biannually) and for-hire vehicles (biennially).

3.8.2 Aerial Imagery Data

We digitize high-resolution, natural color aerial orthoimagery captured over New York City between 2010 and 2017. Our sample of 29 scenes comes from state and federal government sources (US Dept. of Agriculture, the New York City government), online services (Google, ESRI, Bing), and a commercial vendor (NearMap Ltd.). We have excluded a second commercial vendor (Eagle View Technologies, Inc.) due to licensing cost. Unless otherwise noted below, the nominal resolution for imagery is 15 cm or better. Figure 3.7 illustrates this variation in resolution. The sources are:

- 3 scenes from National Agriculture Imagery Program of the US Department of Agriculture (see fsa.usda.gov/programs-and-services /aerial- photography/imagery-

programs/naip-imagery for sources). July 10, 2011 (nominal 1m resolution). June 22, 2013 (nominal 1m resolution). May 22, 2015 (nominal 50cm resolution).

- 4 scenes from NY State Digital Orthoimagery Program, distributed by NYC government at maps.nyc.gov/tiles.
- 11 scenes from Google aerial imagery, obtained through the Google Earth Pro desktop application.
- 1 scene from Bing Maps Aerial Imagery layer.
- 1 scene from ESRI World Imagery layer. Resolution is unknown, but is provided at approximately 30 cm.
- 9 scenes from Nearmap Ltd. (© 2017).

On every image we digitize vehicle locations in two areas of interest: i) an area in the northeast of Manhattan that includes parts of the Upper East Side and East Harlem neighborhoods, delimited between 1st and 5th avenues and between East 79th and East 128th streets, and ii) an area in the eastern Midtown district, delimited between 1st and 6th avenues and between 38th and 57th streets (we select the same area as Zhan et al. (2013), which includes the densest zone in midtown Manhattan). We manually tag five classes of vehicles: yellow taxis, green taxis, black cars, cars of any color other than black and trucks, excluding city buses. Yellow and green taxis are tagged anywhere on the roadbed, whether parked or in apparent motion. Cars and trucks are tagged only on traffic lanes. We exclude taxis on parking lots off the roadbed.

A fraction of our aerial images are subject to some degree of road occlusion, i.e. portions of the roadbed are obstructed from view by tall buildings due to the angle at which the camera captured the aerial photography. We impute observations that are missing due to visual occlusion with the following procedure:

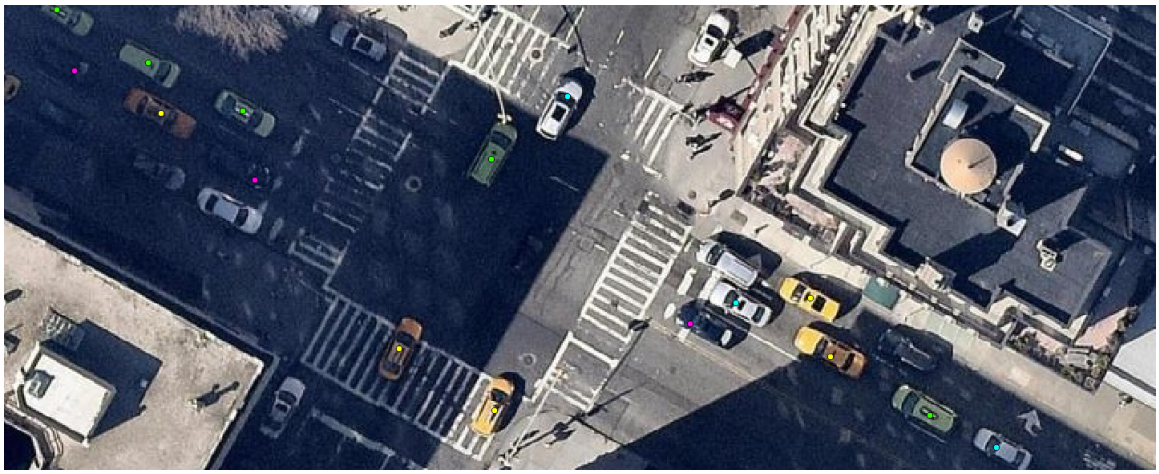
1. For any aerial scene i we start with the set of all observed and digitized vehicle locations L_i. We refer to any aerial scene with some occlusion as a *recipient* image, and any scene with no occlusions as a *donor* image. We separately process the images that were captured in the period before or after August 2013. For donors, we refer to the sets of these images as $\mathcal{D}_b$ and $\mathcal{D}_a$, respectively.

2. For every recipient, we digitize a layer of polygons over every occluded road surface, and refer to the set of these areas as O_r.

Figure 3.7: Aerial Imagery Examples. Madison Ave and East 96th Street

(a) Source: National Agricultural Imagery Program. July 10, 2011. Time estimated as 12.30 pm.

(b) Source: Nearmap Ltd. (© 2017). April 12, 2015, within 5 minutes of 2.57 pm.



Note: the example areas displayed have an area of 0.7 acres, and are intended to illustrate variation in image resolution (1 meter vs. 15 cm). Our area of interest consists of 1087 acres (1.7 square miles) on the east side of northern Manhattan, delimited by 1st Ave and 5th Ave and between East 82nd Street and East 125th Street. Circle icons on the aerial imagery indicate vehicles labeled as taxis (yellow), boro taxis (green), cars (cyan) or black cars (magenta).

3. For every donor d, and for every recipient r that is in the same period $t = \{b, a\}$ as d, we partition L_d into $\tilde{L}_d^r$, the set of vehicle locations that would be counterfactually occluded under the occlusion layer O_r, and $\tilde{L}_d^{-r}$, the set of vehicles that would remain visible under the counterfactual occlusion layer O_r.

4. For every recipient r in period $t = \{b, a\}$, we assemble a donor pool of every observed but counterfactually occluded vehicle location $P_r = \bigcup_{d \in \mathcal{D}_t} \tilde{L}_d^r$.

5. We define the augmented vehicle location set $L_r^+ = L_r \cup P_r$, where to each element in L_r we attach a weight of 1, and to each element in P_r from donor d we assign the scalar weight $w_{dr} = (1/(|\mathcal{D}_t|) \left[|L_r|/(|L_d| - |\tilde{L}_d^r|) \right]$.

The first factor in the weight w_{dr} simply averages over the set of donors. The second maintains proportionality between vehicle counts in the image areas that are mutually observed in the recipient and each donor image. For example, if donor d contains half as many vehicles as a recipient r in the area that is mutually observable in both images, any vehicle locations donated from d to r will be weighted by 1/2, and further re-weighted in proportion to the number of donor images. We apply the above procedure separately for each area of interest and vehicle type.

Buildings in area (i) are relatively low and imputations have little impact on counts and densities: although 16 of 29 images contain some imputations, imputations only increase the total vehicle count in this area by 1.1% and additions to the sample are only substantial in four images: the five largest percentage increases in weighted vehicle count are 11.3%, 10.1%, 7.8%, 7.8% and 1.2%. Area (ii) in midtown Manhattan is dense with skyscrapers, so even small deviations from a perpendicular angle in aerial photography can result in road occlusion. The imputation procedure described above increases the total vehicle count by 6.1% in this area, with imputed vehicles added to 22 scenes. The vehicle count is increased by more than 10% in 8 scenes, and between 10 and 1% in a further 9 scenes. Details on raw, imputed and total counts for both areas are provided in Tables 3.11 and 3.12 below.

Table 3.11: Aerial Imagery Sources and Counts. Upper East Side and East Harlem

Date	Source	Raw count	Imputed	Total	Time of day
6/15/2010	NYC Ortho	2,819		2,819.0	10:30 am
6/17/2010	Google Earth	2,314	181.2	2,495.2	10:15 am
3/26/2011	ESRI Aerial	1,649	2.5	1,651.5	12:30 pm
6/2/2011	Google Earth	3,279		3,279.0	11:45 am
7/10/2011	NAIP	1,920	3.7	1,923.7	12:30 pm
3/10/2012	Google Earth	1,708		1,708.0	11:30 am
6/15/2012 to 6/20/2012	NYC Ortho	3,271		3,271.0	11:15 am
8/6/2012	Google Earth	2,786		2,786.0	9:30 am
11/5/2012	Google Earth	2,423	274.4	2,697.4	10:30 am
5/26/2013	Google Earth	1,257		1,257.0	11:00 am
6/22/2013	NAIP	1,297	100.6	1,397.6	8:30 am
6/1/2014	NYC Ortho	2,160	27.1	2,187.1	11:00 am
6/19/2014	Google Earth	3,120	11.8	3,131.8	12:30 pm
7/6/2014 to 8/10/2014	Bing Aerial	2,657		2,657.0	9:15 am
9/15/2014	Nearmap	2,893		2,893.0	9:40 am to 10:08 am
10/11/2014	Google Earth	1,150	116.4	1,266.4	10:45 am
4/12/2015	Nearmap	2,686	3.9	2,689.9	2:44 pm to 3:11 pm
5/22/2015	NAIP	2,770	2.3	2,772.3	9:45 am
6/22/2015	Nearmap	3,016		3,016.0	10:15 am to 10:47 am
9/6/2015	Google Earth	1,223	0.2	1,223.2	11:30 am
9/6/2015	Nearmap	1,884		1,884.0	1:01 pm to 1:30 pm
10/26/2015	Nearmap	2,574		2,574.0	11:35 am to 12:03 pm
3/26/2016 to 4/5/2016	NYC Ortho	3,750		3,750.0	12:45 pm
4/16/2016	Nearmap	2,534	14.3	2,548.3	11:31 am to 12:09 pm
6/24/2016	Google Earth	2,123	0.3	2,123.3	11:30 am
6/25/2016	Google Earth	1,617	1.4	1,618.4	11:30 am
7/16/2016	Nearmap	1,876		1,876.0	11:11 am to 11:37 am
10/15/2016	Nearmap	2,667	7.0	2,674.0	11:34 am to 11:58 am
4/9/2017	Nearmap	2,632	3.8	2,635.8	1:05 pm to 1:27 pm

Note: Capture dates are provided in metadata for all sources, except two NYC scenes for which a date range is provided, and the Bing Aerial layer which was captured between July 6 and August 10, 2014 according to imagery service API metadata. Time-of-day of image capture is provided in a precise range by Nearmap. For other sources, the time-of-day is estimated by using the shade of the Columbus Circle monument as a sundial (calendar-adjusted). Building shadows in area of interest are consistent with monument shadows. Nearmap metadata and imagery validates our sundial method to within half-hour accuracy.

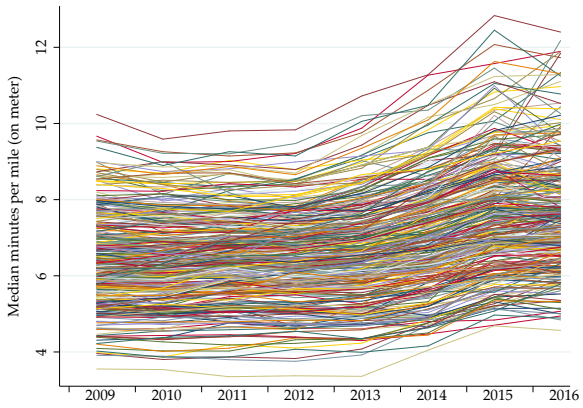
Table 3.12: Aerial Imagery Sources and Counts. Midtown

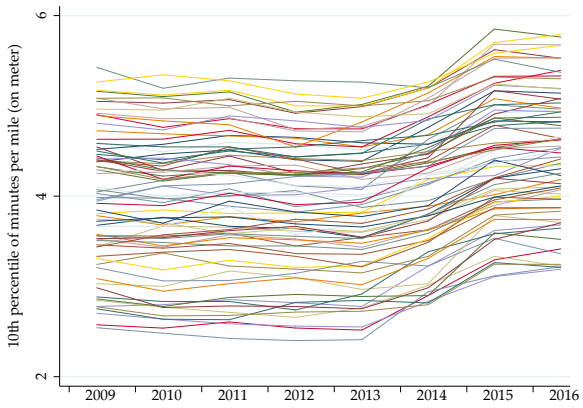
Date	Source	Raw count	Imputed	Total	Time of day
6/15/2010	NYC Ortho	3,162		3,162.0	10:30 am
6/17/2010	Google Earth	1,815	463.8	2,278.8	10:15 am
3/26/2011	ESRI Aerial	1,032	556.2	1,588.2	12:30 pm
6/2/2011	Google Earth	2,886	17.3	2,903.3	11:45 am
7/10/2011	NAIP	1,930		1,930.0	12:30 pm
3/10/2012	Google Earth	1,589	7.4	1,596.4	11:30 am
6/15/2012 to 6/20/2012	NYC Ortho	3,590		3,590.0	11:15 am
8/6/2012	Google Earth	3,150	65.6	3,215.6	9:30 am
11/5/2012	Google Earth	1,430	701.2	2,131.2	10:30 am
5/26/2013	Google Earth	1,110	10.9	1,120.9	11:00 am
6/22/2013	NAIP	828	296.1	1,124.1	8:30 am
6/1/2014	NYC Ortho	1,929	194.0	2,123.0	11:00 am
6/19/2014	Google Earth	2,497	21.2	2,518.2	12:30 pm
7/6/2014 to 8/10/2014	Bing Aerial	2,167		2,167.0	9:15 am
9/15/2014	Nearmap	3,557		3,557.0	8:34 am to 10:09 am
10/11/2014	Google Earth	1,017	222.1	1,239.1	10:45 am
4/12/2015	Nearmap	2,361	55.3	2,411.3	3:11 pm to 3:30 pm
5/22/2015	NAIP	2,817	303.3	3,120.3	9:45 am
6/22/2015	Nearmap	3,137		3,137.0	10:10 am to 10:38 am
9/6/2015	Google Earth	987	14.3	1,001.3	11:30 am
9/6/2015	Nearmap	1,989	174.9	2,163.9	12:43 pm to 1:01 pm
10/26/2015	Nearmap	2,229	94.8	2,323.8	11:15 am to 11:34 am
3/26/2016 to 4/5/2016	NYC Ortho	3,867		3,867.0	12:45 pm
4/16/2016	Nearmap	1,965	114.8	2,079.8	12:00 to 12:18 pm
6/24/2016	Google Earth	2,270	109.9	2,379.9	11:30 am
6/25/2016	Google Earth	2,132	221.6	2,353.6	11:30 am
7/16/2016	Nearmap	1,879	50.5	1,929.5	11:04 am to 11:36 am
10/15/2016	Nearmap	2,437	111.1	2,548.1	11:18 am to 11:26 am
4/9/2017	Nearmap	2,107	82.5	2,189.5	12:56 pm to 1:04 pm

3.8.3 Evidence of the Slowdown in New York City Streets and Highways

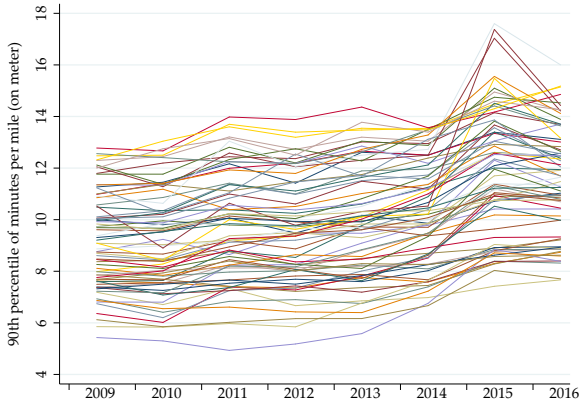
Figure 3.8: Travel Times (percentiles of minutes per mile) for Origin-Destination Pairs in Midtown Manhattan, Month of June 2009 to 2016

(a) Medians, 342 pairs between 19 midtown

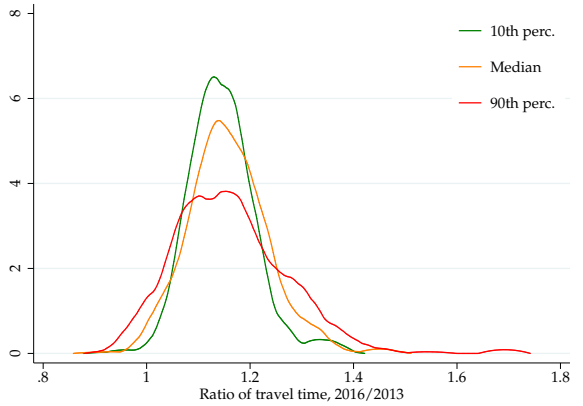




(c) 90th perc., 72 pairs to 4 zones on East River



(d) Densities over 342 pairs for June 2016/June 2013 change ratios

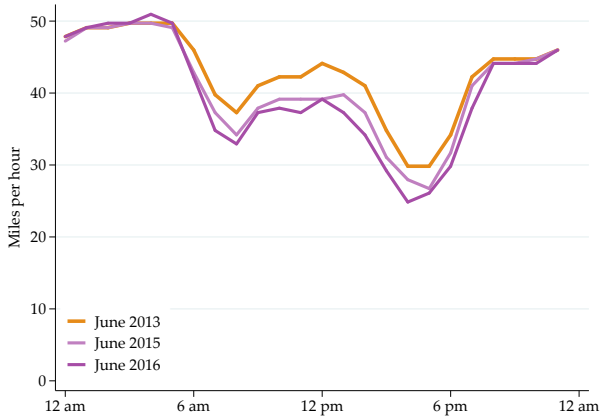


Note: All figures are constructed from raw coordinates on TPEP trip data matched to "taxi zones," during the month of June from 2009 to 2016. The measures plotted are quantiles of the minutes per mile on the taxi meter. "Taxi zones" are a neighborhood-level unit of spatial aggregation defined by the TLC and at which for-hire-vehicle bases (e.g. livery, ridehail) are required to report trip originations. Midtown Manhattan between 14th and 59th streets contains 19 taxi zones. Densities in Figure 3.8d are weighted by trips in June 2013.

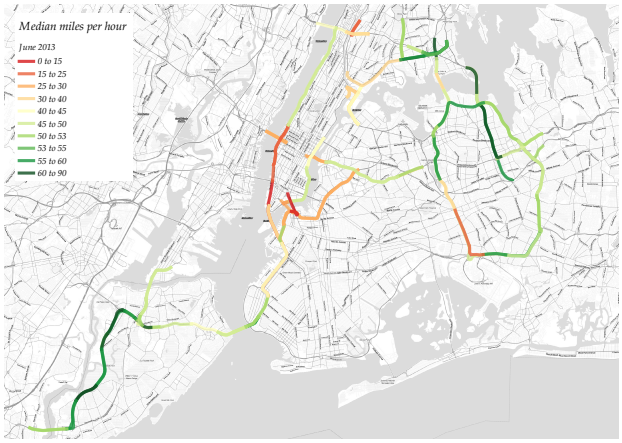
Figure 3.9: Historical Speeds, Archived from Live EZ-Pass Traffic Sensor Feeds

(a) Speed deciles, by month and across links

(b) Median speed, by hour and across links, selected months



(c) Median link speed for June 2013



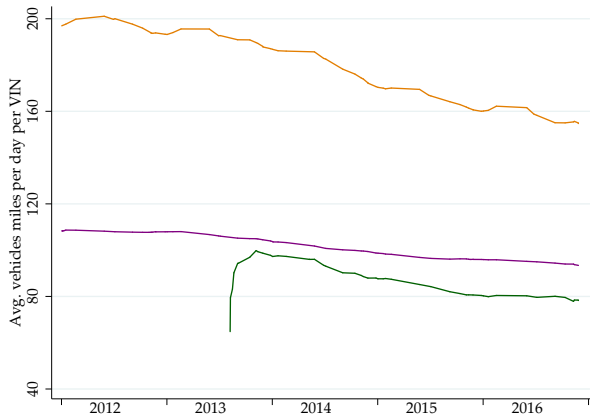
(d) Median link speed for June 2016

Note: Speeds are obtained from archived feeds of live sensor data from the NY State EZ-Pass traffic monitoring system. Sources are described in 3.8.1. Fig. (a) plots the time series for deciles of all EZ-pass traffic sensor speed measurements (10th to 90th percentile, in miles per hour) and across all links in a given month (subselections of this data, such as weekdays during daytime, display similar patterns). Percentiles are plotted as lines from May 2013 through August 2013 and from April 2015 onwards, which are the months for which continuously archived data is available. Black circles plot the median speed across links for single snapshots of the data that were archived by archive.org, for nine additional months in which some of this data is available. Fig. (b) plots median speeds across links per hour for the month of June in 2013, 2015 and 2016. Figures (c) and (d) map the median speeds at the link level for June 2013 and June 2016, respectively. On both maps the speed levels are displayed on a color ramp that corresponds approximately to deciles of the speed distribution in June 2013, as plotted in the time series shown in fig. (a). The network of road segments linked through EZ-pass sensors has extended over time, the maps in figures (c) and (d) plot the 115 segments that were active in June 2013 and June 2016. Links contain overlapping segments, which are shown by means of transparencies.

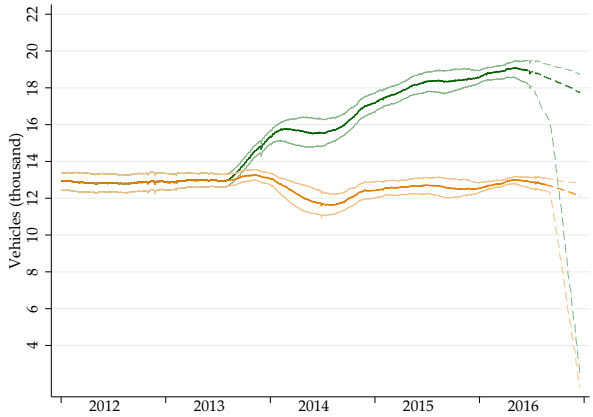
Figure 3.10: Vehicle Miles Traveled (VMTs) by Vehicle Class, Estimated from Odometer Data

Note: methodology is described in 3.8.5. In particular, note that plotted bounds arise from limiting methodological assumptions, and are not confidence intervals. Figure (a) plots the daily average of interpolated miles per day per VIN for each class of vehicles, with orange for taxis, green for boro taxis and purple for all for-hire vehicles. Miles per day per VIN require two odometer inspections around the interpolation date. Figures (b)-(c) plot the estimated stock of registered vehicles per day, and figures (d)-(e) plot an estimate of vehicle-miles traveled per day, i.e. the product of the vehicle stock times the daily average flow of miles per day per VIN for each class. In figures (b) and (d) the color yellow indicates yellow taxis, whereas green indicates yellow plus boro taxis.

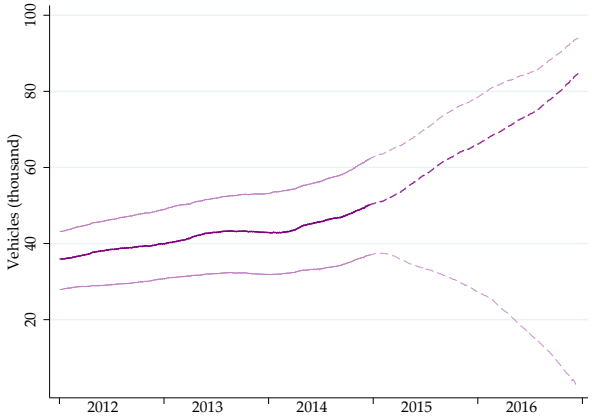
(a) Avg. miles per day per VIN and class interpolated from odometer readings



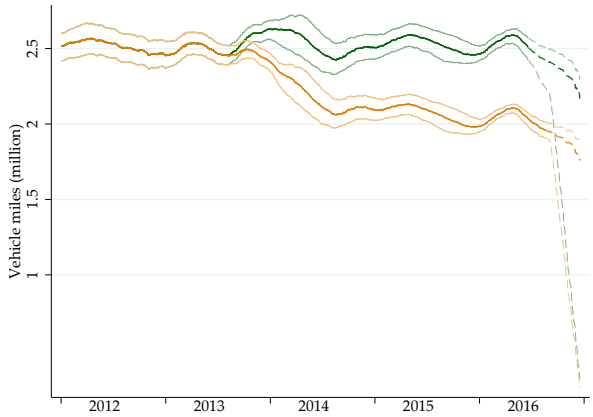
(b) Stock of registered yellow and boro taxis



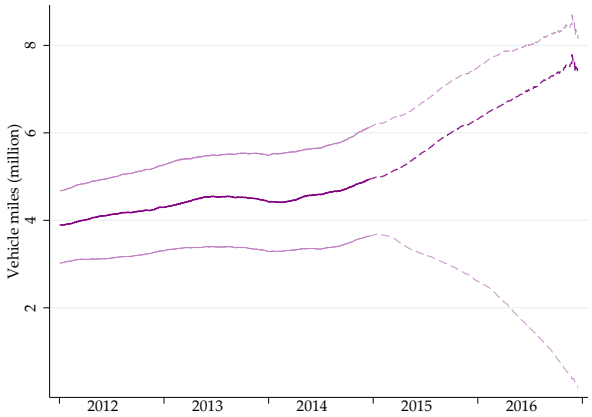
(c) Stock of registered for-hire vehicles



(d) Yellow and boro taxi total VMTs per day



(e) For-hire vehicle total VMTs per day



3.8.4 GPS Data Validation

This paper employs geographic coordinates of taxi pickups and drop-offs recorded by GPS units onboard taxis (TPEP units) to construct historical speed measures and to track the location of taxi activity. In this appendix, we evaluate the robustness of our identification strategy (in particular, our reduced form estimates) to measurement error in these geographic coordinates. Although there is no obvious direction of bias, GPS accuracy in our area and period of study was subject to spatial and time trends that could, in principle, act as a confounder.

The accuracy of navigation satellite systems depends on the number of signal-emitting satellites within line-of-sight of the receiver, and tall buildings may create "urban canyons" that reduce the number of satellites in view. In northern Manhattan the height and density of buildings increases in the downtown direction. Since the height of buildings surrounding a road segment rarely changes, GPS accuracy could be a time-invariant attribute of road segments that can be controlled for with road segment fixed effects (i.e. run-bin f.e. in the notation used in the paper). However, the number of "GPS satellites" that at any given moment are directly overhead and available to receivers in New York City has increased substantially in the past decade with the deployment of satellite systems from Russia, the European Union and China.[51] If trends in GPS satellite availability have a differential impact on coordinate accuracy that depends on the heights of surrounding buildings, GPS accuracy could follow differential trends across the boro and "hail-exclusion" zones.

We employ two measures of potential satellite obstruction throughout northern Manhattan, illustrated in Figure 3.11: the average number of floors of the buildings that overlap a road segment in the area of interest (floor numbers are color-coded in blue), and a measurement of sky-view factor (SVF), which is the share of sky in a hemispherical field of vision from the ground. We employ SVF measures from Liang et al. (2017), which were computed from Google Street View imagery at the midpoint of every city block in Manhattan. Figure 3.11 illustrates that building height (and sky-view-factor) are correlated with latitude in northern Manhattan.[52] Table 3.13 estimates the baseline reduced form result from Table 3.5, Col. 2 while including controls for trends in GPS accuracy. Col. 1

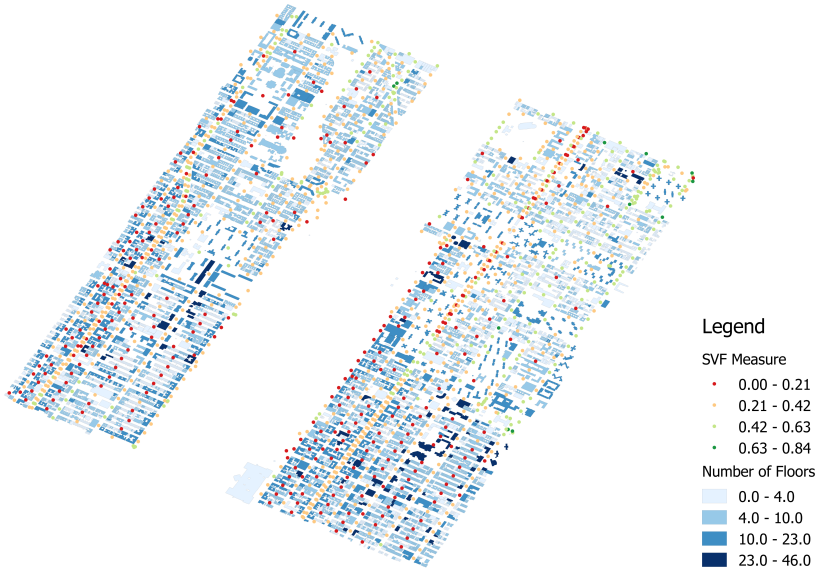
[51] GPS is a name that is specific to the United States government's global navigation satellite system, which was fully deployed to its current constellation size (32 satellites) and coverage level prior to 2009. The satellite constellations of Russia (GLONASS), the EU (Galileo) and China (BeiDou2) have been expanded between January 2009 and June 2016, with net gains of 7, 9 and 19 satellites respectively. The satellites of these systems emit signals that are available to consumer "GPS" units, and the expansion has allowed for significant accuracy improvements worldwide.

[52] Park Avenue is an exception, as it is separated by elevated tracks north of East 101st Street.

includes the average number of completed floors for the buildings with a footprint that overlaps each 10 meter segment (i.e. bin) on every avenue. This control has no impact on the coefficient of interest, which is not surprising since it contains almost no variation over time and all of its cross-sectional variation was previously absorbed by run-bin fixed effects. Col 2. interacts quartiles of the previous measure with linear trends, to allow for arbitrary differential linear trends over ranges of building heights. Col. 3 interacts linear trends with quartiles of street-view-factors, where values for bins in which Liang et al. (2017) did not calculate a factor are interpolated from neighboring measures: the estimated coefficient barely changes. Results from Cols. 2 and 3 suggest that any differential trends in GPS accuracy by level of urban density did not have an impact on the measurement of street-speed that correlates in space and time with the roll-out of the boro program.

We next employ a direct measure of taxi GPS accuracy: we filter all the taxi pickup or drop-off coordinates in our area of interest through the building footprints in the area. Although coordinates may be in error even if recorded at a street curb, one type of error that we can measure over space and time is the share of trip endpoints that were placed by the taxi's GPS unit within the walls of an existing structure. Using this definition of GPS accuracy we do find that accuracy evolved differently over the treatment and control zones: 8.8% of trip endpoints in the hail-exclusion zone, and 5.4% in the boro zone were located within a building footprint in the 56 months leading up to August 2013, whereas fail rates declined to 7.2% and 2.6%, respectively, in the 34 months from September 2013 onward. Apparently, the deployment of additional satellites has had a larger effect on the accuracy of GPS fixes at an intermediate rather than at the higher end of urban density. Columns 4 and 5 include levels and logs of this GPS failure rate. Once again, our baseline estimate of the impact of the boro program is not sensitive to the inclusion of this control for GPS accuracy.

Figure 3.11: Building Footprints, Number of Floors and Sky-View Factors



Note: Sky-view factor (SVF) is the share of the vertical field of vision that is composed of sky. SVF measures were computed by Liang et al. (2017) from Google Street View imagery. Building footprints and number of floors are from data.cityofnewyork.us/Housing-Development/Building-Footprints/nqwf-w8eh.

Table 3.13: Robustness of Reduced Form Impact to Measurement Error in GPS Units

	(1)	(2)	(3)	(4)	(5)
	Log avg. s/m	Log avg. s/m	Log avg. s/m	Log avg. s/m	Log avg. s/m
BZ x Roll-out	0.081	0.081	0.078	0.081	0.081
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Avg. number of floors	0.000				
	(0.000)				
Avg. num. floors: quartiles × linear trends		Y			
Sky view factor: quartiles × linear trends			Y		
GPS fail rate				0.008	
				(0.002)	
Log GPS fail rate					0.000
					(0.000)
Standard controls	Y	Y	Y	Y	Y
Run-bin × Run-my FE	Y	Y	Y	Y	Y
Observations	406260	406260	402750	406260	406260
R^2	0.97	0.97	0.97	0.97	0.97
R^2-within	0.15	0.15	0.15	0.15	0.15

Note: Unless otherwise specified, robust standard errors are two-way clustered at the level of run× bin and run× month. In Col. 3, street-view-factors are interpolated from neighboring measures for bins without data from Liang et al. (2017), and 39 bins are dropped from the sample due to lack of data for interpolation. The estimated coefficient on the subsample of 1,098 bins for which SVFs are available is almost unchanged. Cols. 2 and 3 include interactions of GPS accuracy determinants with linear trends; interactions with the *post* dummy or the *Boro roll-out* trend also have no effect on the coefficient of interest.

3.8.5 Method for Estimating Vehicle Miles Traveled from Odometer Inspection Data

In this appendix we explain our methodology for calculating the total number of vehicle-miles traveled (VMTs) per day for each of the following classes of vehicles: yellow taxis, boro taxis and for-hire vehicles (FHVs), the latter a class that includes traditional black cars and other car services as well as, more recently, ridehail providers. We employ odometer readings per vehicle identification number (VIN) from 2010 through 2016, measured at regular inspections required by the Vehicle Inspection Program of New York City's Taxi and Limousine Commission and obtained through freedom-of-information requests.

Vehicles in each class are required to undergo regular inspections: within 3 months for yellow taxis, 6 months for boro taxis and 2 years for FHVs. The empirical challenge to estimating time series of VMTs from regular odometer inspections is that the data is naturally censored. Vehicles require an inspection to enter car service, but may exit car service at any point after an inspection, and this exit will not be observed until the vehicle fails to show up for an inspection within the required window. To address this missing data problem, we separately estimate daily flows of VMTs from the daily stock of vehicles on the road. Estimating average mile flows conditional on the set of vehicles that are confirmed survivors (i.e. are between inspections) is straightforward: we interpolate the odometer readings for each survivor to obtain daily flows, and average these daily flows over all survivors.

Accounting for censoring is more involved. We cannot estimate the timing of exit because we do not observe exit dates on any part of our sample; we only observe whether vehicles fail to show within the re-inspection window. Additionally, censoring interacts with the end date of our inspection sample (December 31, 2016) to create a period of uncertain exit. We use FHVs as an example hereafter but employ the same approach for each class of vehicles: whereas we can be certain that a VIN that was last seen prior to December 31, 2014 has exited FHV service some time over the following two years, any VIN that was last observed from January 1, 2015 onwards may have exited or may still be in service as of December 31, 2016.

Using the subsample of VIN-inspection dates for which a subsequent re-inspection or exit has occurred with certainty, we estimate the 2-year survival rate for each interval of 50,000 miles on the odometer (through 400K miles, all higher mileages are binned together). We then compute the stock of vehicles on the road as of date t as the sum of confirmed survivors (i.e. the VIN is re-inspected after t) plus an estimate of expected

survivors, which consist of expected confirmed exiters (since exit date is not identified, we assume it is equiprobable over the inspection window) and expected uncertain exiters (we assume the constant hazard rate that accumulates to the estimated 2-year survival rate). We then compute our baseline daily VMT estimates as the product of daily average VMT flows per VIN and the expected stock of active VINs.

Out of necessity, our approach makes two strong assumptions: equiprobable daily exit rates for confirmed exiters and constant hazard rates for uncertain exiters. As bounds on the previous approach, we consider two alternative assumptions: we obtain an upper bound by assuming that all vehicles survive until we observe them to miss a re-inspection date, and a lower bound by assuming exit immediately after the last inspection. Note that the lower bound approach mechanically drives down the stock of estimated cars during the uncertain exit window, all the way down to zero on Jan 1, 2017.

Our estimate of daily VMT flows per vehicle, as well as baseline and bound estimates for stocks and total VMT flows are plotted in Figure 3.10. Note that shorter re-inspection windows for yellow taxis result in tighter estimates and shorter uncertain exit periods. We start our plots on January 1, 2012 since a symmetric source of uncertainty over entry occurs at the start of our sample period. We do not allocate FHVs miles between ridehail and traditional car services. We do not believe that this is possible without de-anonymized, VIN level data from ridehail providers because vehicles affiliated to traditional car service bases are supplied on ridehail platforms. Throughout our application we employ the sum of both modes as the unit of analysis.

3.8.6 Additional Figures and Tables

Table 3.14: Wald Estimates of the Congestion Elasticity of Taxi Pickups. Spline Speeds

	(1)	(2)	(3)	(4)	(5)	(6)
	Log Bsp. s/m	Log Bsp. s/m	Log Bsp. s/m	Log Bsp. s/m	Log Bsp. s/m	Log Bsp. s/m
	(OLS)	(2SLS)	(OLS)	(2SLS)	(OLS)	(2SLS)
Log pickups	-0.006	0.132	-0.005	0.188	0.002	0.038
	(0.006)	(0.027)	(0.007)	(0.038)	(0.007)	(0.023)
Bikelane	-0.071	-0.106	-0.183	-0.271	0.122	0.125
	(0.052)	(0.056)	(0.077)	(0.087)	(0.041)	(0.041)
Citibike	0.062	0.115	0.080	0.156	0.036	0.049
	(0.035)	(0.039)	(0.049)	(0.055)	(0.016)	(0.020)
311 complaints	0.003	0.002	0.004	0.003	0.001	0.001
	(0.003)	(0.003)	(0.005)	(0.005)	(0.002)	(0.002)
Potholes	-0.005	-0.007	-0.033	-0.037	0.023	0.023
	(0.009)	(0.010)	(0.017)	(0.018)	(0.007)	(0.007)
Run-bin and Run-m-y FE	Y	Y	Y	Y	Y	Y
Sample	All	All	East side	East side	West side	West side
Observations	406260	406260	216000	216000	190260	190260
R^2	0.64	0.63	0.54	0.53	0.74	0.74

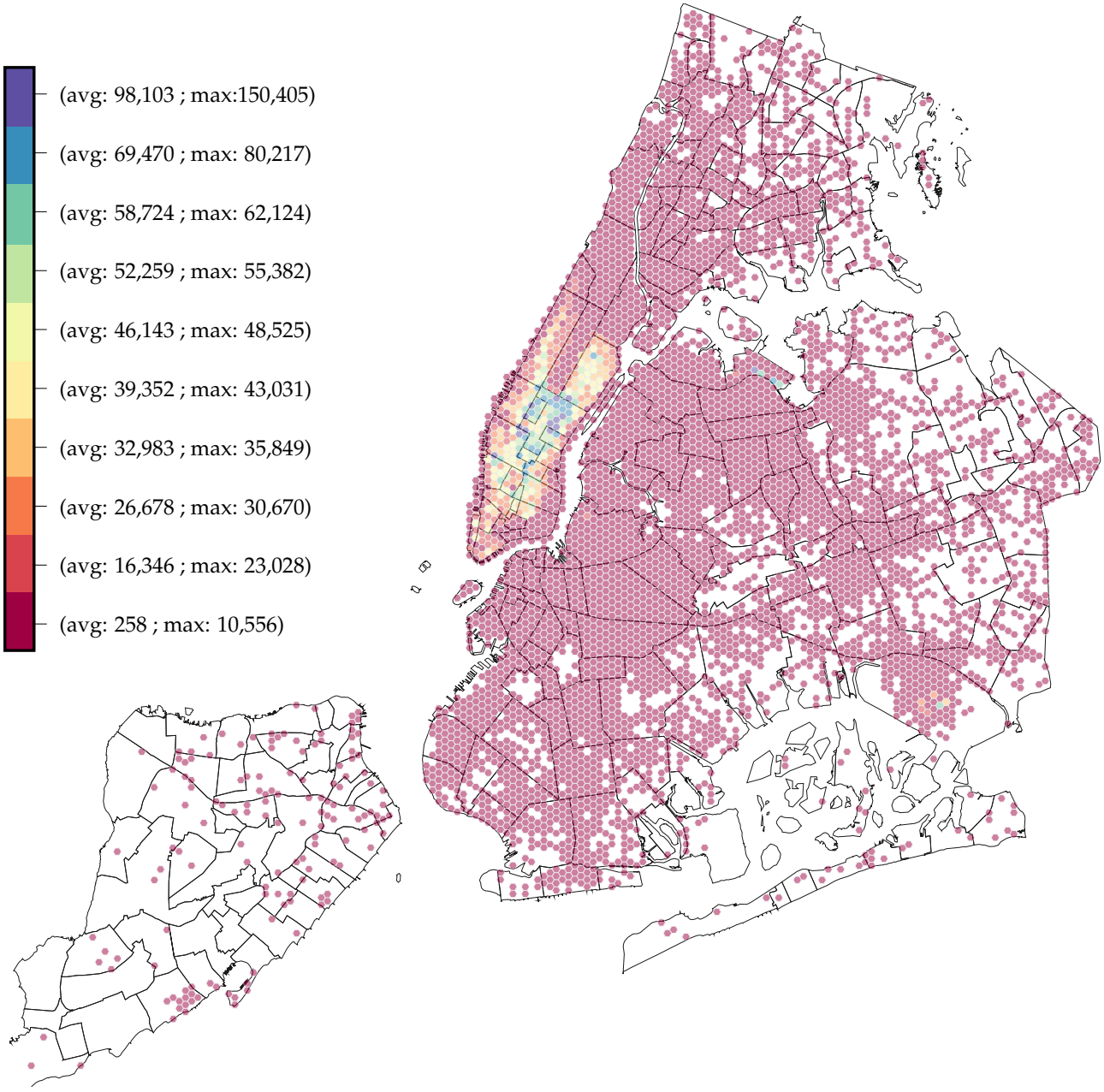
Note: Unless otherwise specified, robust standard errors are two-way clustered at the level of run× bin and run× month. Weak instrument tests are not reported, but the lowest F-stat in the table (for column 9, the West side, Bspline specification) is 35,315. About 1.16% of the 10-meter avenue segments had no taxi pickups within a particular month: we set the dependent variable at log(pickups + 0.1).

Table 3.15: Program Impact Coefficient by Time Slice

	24 hrs.	6am-8pm	6am-10am	10am-2pm	2pm-4pm	4pm-8pm	8pm-11pm	11pm-6am
Full week coef.	0.081	0.080	0.067	0.091	0.052	0.094	0.079	0.080
Std. err.	(0.004)	(0.004)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)	(0.007)
Avg. trips per run-month-bin	1824.4	1391.4	286.2	398.2	230.1	477.0	236.4	196.6
Weekday coef.	0.077	0.074	0.061	0.081	0.043	0.097	0.076	0.065
Std. err.	(0.005)	(0.005)	(0.007)	(0.006)	(0.007)	(0.005)	(0.006)	(0.007)
Avg. trips per run-month-bin	1306.3	1019.5	225.8	268.8	167.9	357.0	171.0	133.5
Weekend coef.	0.071	0.068	0.045	0.072	0.061	0.073	0.091	0.104
Std. err.	(0.005)	(0.005)	(0.007)	(0.006)	(0.008)	(0.007)	(0.008)	(0.008)
Avg. trips per run-month-bin	518.2	371.9	60.4	129.3	62.2	120.1	65.4	120.0

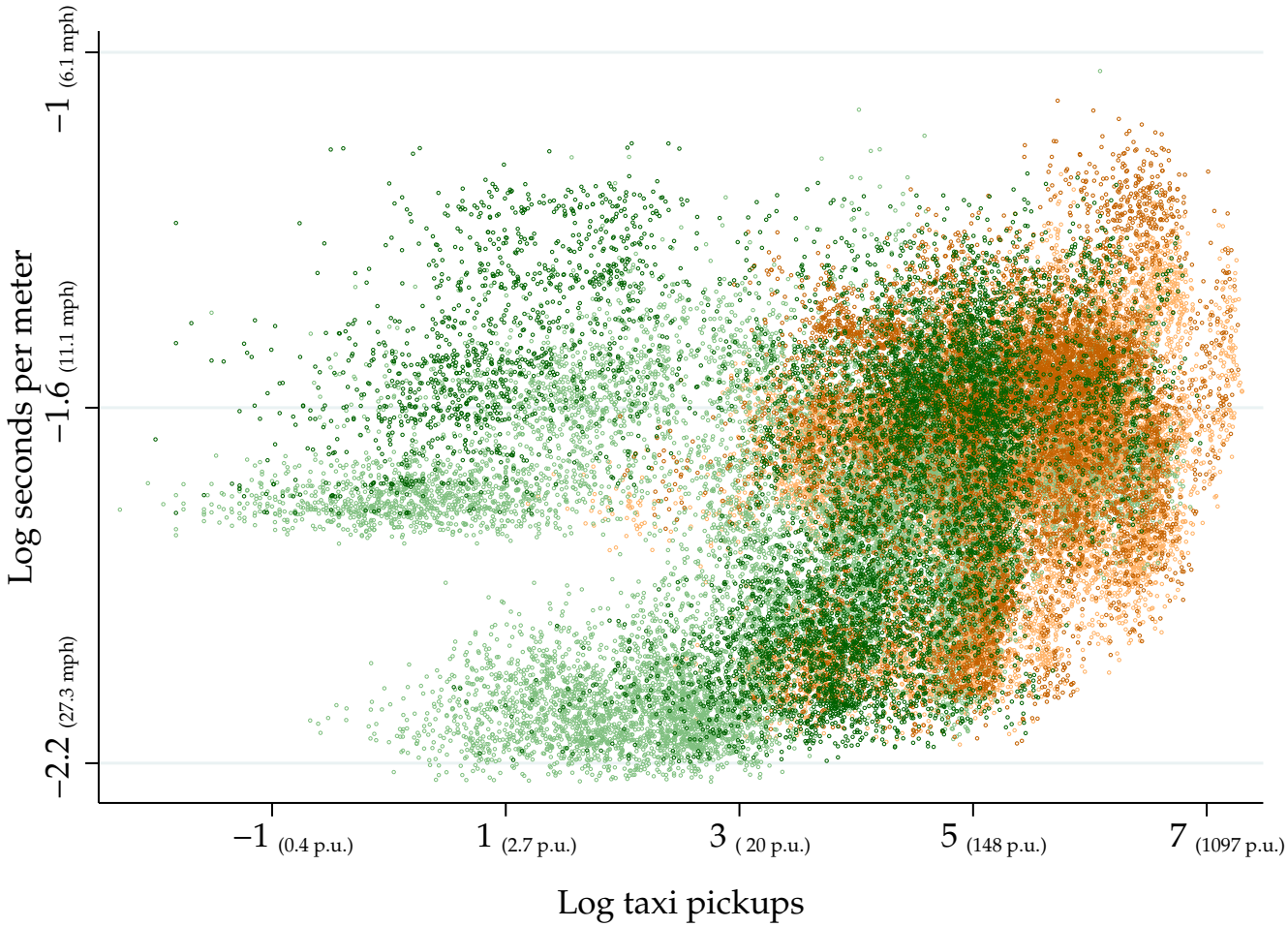
Note: This table reports variations on the the baseline program impact regression (i.e. Column 2 of Table 3.6) in which the average second-per-meter outcome measure is computed under alternative time slices of the within-run trip data sample. The table is ordered in rows for day-of-week and columns for time-of-day slices, and the top left cell is the same specification as Table 3.6, Col. 2. Each cell reports estimated coefficient and standard errors for the boro-zone times roll-out interaction, as well as the average number of trips used in computation of the average travel time for each run, month and 10 meter bin.

Figure 3.12: Spatial Distribution of Yellow Medallion Taxi Pickups. June 2013.



Note: Raw taxi records (TPEP data, described in 3.8.1) for all pickups within New York City during June 2013, binned into 12,004 hexagonal cells (each with an area of 0.081 sq. km, or 19.8 acres). Every color shade accounts for approximately one tenth (1.4 million) of all taxi pickups. For example, the 14 hexagonal cells in purple accounted for 9.73% of all pickups in June 2013, with an average 98,103 pickups per cell, a volume of pickups that is equal to that of the area shaded in red (average 258 pickups per cell) and white (no pickups). Numbers on the color bar indicate the average and maximum number of pickups over the cells in a given color.

Figure 3.13: Traffic Speeds and Taxi Pickups



Note: Sample from Table 3.9, Cols. 1-3. Treatment status and period is indicated by color and shade, according to the following scheme: bins in the treatment "boro" zone are green, control are orange; pre period in a light shade and post period in a dark shade. Each point is the average for both variables of 10 contiguous bins on a given run and month. Axes are in log travel time and log pickups, but absolute values for speed and pickups are included for convenience.

Bibliography

Aaron, Kat. 2015. "If Not Uber, Who's Responsible for This Traffic?" *WNYC Broadcast. August 11, 2015* .

Abraham, Katharine G, Emel Filiz-Ozbay, Erkut Y Ozbay, and Lesley J Turner. 2018. "Framing effects, earnings expectations, and the design of student loan repayment schemes." Tech. rep., National Bureau of Economic Research.

Ahlfeldt, Gabriel M., Stephen J. Redding, Daniel M. Sturm, and Nikolaus Wolf. 2015. "The Economics of Density: Evidence From the Berlin Wall." *Econometrica* 83 (6):2127–2189.

Akbar, Prottoy and Gilles Duranton. 2017. "Measuring the Cost of Congestion in Highly Congested City: Bogotá."

Alaimo, Katherine, Christine M Olson, and Edward A Frongillo. 2001. "Food insufficiency and American school-aged children's cognitive, academic, and psychosocial development." *Pediatrics* 108 (1):44–53.

Anderson, Drew M, Johnathan G Conzelmann, and T Austin Lacy. 2018. "The state of financial knowledge in college: New evidence from a national survey." *Rand Corporation, July* .

Anderson, Michael L. 2014. "Subways, Strikes, and Slowdowns: The Impacts of Public Transit on Traffic Congestion †." *American Economic Review* 104 (9):2763–2796.

Anderson, Michael L., Fangwen Lu, Yiran Zhang, Jun Yang, and Ping Qin. 2016. "Superstitions, Street Traffic, and Subjective Well-Being." *Journal of Public Economics* 142 (Supplement C):1–10.

Angrist, Joshua D., Sydnee Caldwell, and Jonathan V. Hall. 2017. "Uber vs. Taxi: A Driver's Eye View." Working Paper 23891, National Bureau of Economic Research.

Avery, Christopher and Sarah Turner. 2012. "Student Loans: Do College Students Borrow Too Much–Or Not Enough?" *Journal of Economic Perspectives* 26 (1):165–92.

Barr, Andrew, Kelli Bird, and Benjamin L Castleman. 2016. "Prompting active choice among high-risk borrowers: Evidence from a student loan counseling experiment." *EdPolicyWorks Working Paper* .

Barrios, Thomas, Rebecca Diamond, Guido W Imbens, and Michal Kolesár. 2012. "Clustering, spatial correlations, and randomization inference." *Journal of the American Statistical Association* 107 (498):578–591.

Barth, Matthew and Kanok Boriboonsomsin. 2008. "Real-World Carbon Dioxide Impacts of Traffic

Congestion." *Transportation Research Record: Journal of the Transportation Research Board* 2058:163–171.

Bartik, Timothy J. 1987. "The estimation of demand parameters in hedonic price models." *Journal of Political Economy* 95 (1):81–88.

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

Bettinger, Eric P and Brent J Evans. 2019. "College Guidance for All: A Randomized Experiment in Pre-College Advising." *Journal of Policy Analysis and Management* .

Bettinger, Eric P, Bridget Terry Long, Philip Oreopoulos, and Lisa Sanbonmatsu. 2012. "The role of application assistance and information in college decisions: Results from the H&R Block FAFSA experiment." *The Quarterly Journal of Economics* 127 (3):1205–1242.

Brown, Meta, John Grigsby, Wilbert van der Klaauw, Jaya Wen, and Basit Zafar. 2016. "Financial education and the debt behavior of the young." *The Review of Financial Studies* 29 (9):2490–2522.

Buchholz, Nicholas. 2017. "Spatial Equilibrium, Search Frictions and Efficient Regulation in the Taxi Industry." Working paper.

Cameron, A Colin and Douglas L Miller. 2015. "A practitioner's guide to cluster-robust inference." *Journal of Human Resources* 50 (2):317–372.

Castillo, Juan Camilo, Daniel T. Knoepfle, and E. Glen Weyl. 2017. "Surge Pricing Solves the Wild Goose Chase." SSRN Scholarly Paper ID 2890666, Social Science Research Network, Rochester, NY.

Castleman, Benjamin and Joshua Goodman. 2018. "Intensive college counseling and the enrollment and persistence of low-income students." *Education Finance and Policy* 13 (1):19–41.

Castleman, Benjamin L, Saul Schwartz, and Sandy Baum. 2015. "Prompts, personalization, and pay-offs: Strategies to improve the design and delivery of college and financial aid information." Tech. rep., Center on Education Policy and Workforce Competitiveness.

Chen, M. Keith, Judith A. Chevalier, Peter E. Rossi, and Emily Oehlsen. 2017. "The Value of Flexible Work: Evidence from Uber Drivers." Working Paper 23296, National Bureau of Economic Research.

Clewlow, Regina R. and Gouri S. Mishra. 2017. "Disruptive Transportation: The Adoption, Utilization, and Impacts of Ride-Hailing in the United States." Research Report UCD-ITS-RR-17-07, University of California, Davis.

Cohen, Peter, Robert Hahn, Jonathan Hall, Steven Levitt, and Robert Metcalfe. 2016. "Using Big Data to Estimate Consumer Surplus: The Case of Uber." Working Paper 22627, National Bureau of Economic Research.

Consumer Reports, National Research Center. 2016. "College Financing Survey." http://article.images.consumerreports.org/prod/content/dam/cro/magazine-articles/2016/August/Consumer%20Reports%202016%20College%20Financing%20Survey%20Public%20Report.pdf. Accessed: 2019-09-20.

Cotti, Chad, John Gordanier, and Orgul Ozturk. 2018. "When does it count? The timing of food stamp receipt and educational performance." *Economics of Education Review* 66:40–50.

Cox, James C, Daniel Kreisman, and Susan Dynarski. 2018. "Designed to fail: Effects of the default option and information complexity on student loan repayment." Tech. rep., National Bureau of Economic Research.

Cramer, Judd and Alan B. Krueger. 2016. "Disruptive Change in the Taxi Business: The Case of Uber." *American Economic Review: Papers & Proceedings* 106 (5):177–182.

Currie, Janet and Reed Walker. 2011. "Traffic Congestion and Infant Health: Evidence from E-ZPass." *American Economic Journal: Applied Economics* 3 (1):65–90.

Donaldson, Dave and Adam Storeygard. 2016. "The View from Above: Applications of Satellite Data in Economics." *Journal of Economic Perspectives* 30 (4):171–198.

Duranton, Gilles and Matthew A Turner. 2011. "The Fundamental Law of Road Congestion: Evidence from US Cities." *American Economic Review* 101 (6):2616–2652.

Dynarski, Susan and Judith Scott-Clayton. 2013. "Financial Aid Policy: Lessons from Research." *The Future of Children* 23 (1):67–91.

Dynarski, Susan M. 2003. "Does aid matter? Measuring the effect of student aid on college attendance and completion." *American Economic Review* 93 (1):279–288.

Dynarski, Susan M and Judith E Scott-Clayton. 2006. "The cost of complexity in federal student aid: lessons from optimal tax theory and behavioral economics." *National Tax Journal* 59 (2):319–357.

Farber, Henry S. 2015. "Why You Can't Find a Taxi in the Rain and Other Labor Supply Lessons from Cab Drivers." *The Quarterly Journal of Economics* 130 (4):1975–2026.

Fernandes, Daniel, John G Lynch Jr, and Richard G Netemeyer. 2014. "Financial literacy, financial education, and downstream financial behaviors." *Management Science* 60 (8):1861–1883.

Figlio, David N and Joshua Winicki. 2005. "Food for thought: the effects of school accountability

plans on school nutrition." *Journal of Public Economics* 89 (2):381–394.

Fitzpatrick, Maria D and Damon Jones. 2016. "Higher Education, Merit Aid Scholarships and Post-Baccalaureate Migration." *Economics of Education Review* 54:155–172.

Fitzsimmons, Emma G. and Winnie Hu. 2017. "The Downside of Ride-Hailing: More New York City Gridlock." *The New York Times. March 6, 2017* .

Frechette, Guillaume R., Alessandro Lizzeri, and Tobias Salz. 2016. "Frictions in a Competitive, Regulated Market Evidence from Taxis." Tech. rep.

Frisvold, David E. 2015. "Nutrition and cognitive achievement: An evaluation of the School Breakfast Program." *Journal of public economics* 124:91–104.

Gao Hodges, Graham Russell. 2012. *Taxi!: A Social History of the New York City Cabdriver*. Baltimore: NYU Press.

Ge, Yanbo, Christopher R. Knittel, Don MacKenzie, and Stephen Zoepf. 2016. "Racial and Gender Discrimination in Transportation Network Companies." Working Paper 22776, National Bureau of Economic Research.

Gibbons, Steve, Henry G. Overman, and Eleonora Patacchini. 2015. "Chapter 3 - Spatial Methods." In *Handbook of Regional and Urban Economics*, *Handbook of Regional and Urban Economics*, vol. 5, edited by J. Vernon Henderson Gilles Duranton and William C. Strange. Elsevier, 115–168.

Glaeser, Edward L., Scott Duke Kominers, Michael Luca, and Nikhil Naik. 2015. "Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life." Working Paper 21778, National Bureau of Economic Research.

Goodman, Joshua. 2019. "The Labor of Division: Returns to Compulsory High School Math Coursework." *Journal of Labor Economics* 37 (4).

Goodman, Joshua, Oded Gurantz, and Jonathan Smith. 2018. "Take two! SAT retaking and college enrollment gaps." Tech. rep., National Bureau of Economic Research.

Greenshields, B. D. 1948. "The Potential Use of Aerial Photographs in Traffic Analysis." In *Proceedings of the Highway Research Board*, vol. 27. 291–297.

Gurantz, Oded, Matea Pender, Zachary Mabel, Cassandra Larson, and Eric Bettinger. 2019. "Virtual Advising for High-Achieving HighSchool Students." Tech. Rep. 19-126, Ed Working Paper Series.

Haggag, Kareem, Brian McManus, and Giovanni Paci. 2017. "Learning by Driving: Productivity Improvements by New York City Taxi Drivers." *American Economic Journal: Applied Economics* 9 (1):70–95.

Haggag, Kareem and Giovanni Paci. 2014. "Default Tips." *American Economic Journal: Applied Economics* 6 (3):1–19.

Hall, Andrew, LN Khanh, TH Son, NQ Dung, RG Lansdown, DT Dar, NT Hanh, Helen Moestue, Ha Huy Khoi, and DA Bundy. 2001. "An association between chronic undernutrition and educational test scores in Vietnamese children." *European Journal of Clinical Nutrition* 55 (9):801–804.

Hall, Jonathan V., John J. Horton, and Daniel T. Knoepfle. 2017. "Labor Market Equilibration: Evidence from Uber."

Hall, Jonathan V. and Alan B. Krueger. 2016. "An Analysis of the Labor Market for Uber's Driver-Partners in the United States." Working Paper 22843, National Bureau of Economic Research.

Hamilton, Timothy L. and Casey J. Wichman. 2017. "Bicycle Infrastructure and Traffic Congestion: Evidence from DC's Capital Bikeshare." *Journal of Environmental Economics and Management* Forthcoming.

Hanna, Rema, Gabriel Kreindler, and Benjamin A. Olken. 2017. "Citywide Effects of High-Occupancy Vehicle Restrictions: Evidence from "Three-in-One" in Jakarta." *Science* 357 (6346):89–93.

Harvey, Melody. 2019. "Impact of Financial Education Mandates on Younger Consumers' Use of Alternative Financial Services." *Journal of Consumer Affairs* 53 (3):731–769.

Hoxby, Caroline M and Sarah Turner. 2015. "What high-achieving low-income students know about college." *American Economic Review* 105 (5):514–17.

Hsieh, Chang-Tai and Enrico Moretti. 2017. "Housing Constraints and Spatial Misallocation."

Huston, Sandra J. 2010. "Measuring financial literacy." *Journal of Consumer Affairs* 44 (2):296–316.

Ivanovic, Daniza, Magaly Vásquez, Marcela Aguayo, Digna Ballester, Maximiliano Marambio, and Isabel Zacarías. 1992. "Nutrition and education. III. Educational achievement and food habits of Chilean elementary and high school graduates." *Archivos latinoamericanos de nutricion* 42 (1):9–14.

Jackson, C. Kirabo and Henry S Schneider. 2011. "Do Social Connections Reduce Moral Hazard? Evidence from the New York City Taxi Industry." *American Economic Journal: Applied Economics* 3 (3):244–267.

Johnson, A. N. 1928. "Maryland Aerial Survey of Highway Traffic Between Baltimore and Washington." In *Proceedings of the Highway Research Board*, vol. 8. 106–115.

Kahneman, Daniel and Alan B. Krueger. 2006. "Developments in the Measurement of Subjective

Well-Being." *The journal of economic perspectives* 20 (1):3–24.

Knittel, Christopher R., Douglas L. Miller, and Nicholas J. Sanders. 2016. "Caution, Drivers! Children Present: Traffic, Pollution, and Infant Health." *The Review of Economics and Statistics* 98 (2):350–366.

Kofoed, Michael S. 2017. "To apply or not to apply: FAFSA completion and financial aid gaps." *Research in Higher Education* 58 (1):1–39.

Kraft, Matthew A. 2018. "Interpreting effect sizes of education interventions." Tech. rep., Brown University Working Paper. Downloaded Tuesday, April 16, 2019, from . . . .

Lagos, Ricardo. 2003. "An Analysis of the Market for Taxicab Rides in New York City." *International Economic Review* 44 (2):423–434.

Li, Shanjun, Avralt-Od Purevjav, and Jun Yang. 2017. "The Marginal Cost of Traffic Congestion and Road Pricing: Evidence from a Natural Experiment in Beijing." SSRN Scholarly Paper ID 2948619, Social Science Research Network, Rochester, NY.

Liang, Jianming, Jianhua Gong, Jun Sun, Jieping Zhou, Wenhang Li, Yi Li, Jin Liu, and Shen Shen. 2017. "Automatic Sky View Factor Estimation from Street View Photographs—A Big Data Approach." *Remote Sensing* 9 (5):411.

Looney, Adam, Constantine Yannelis et al. 2019. "The Consequences of Student Loan Credit Expansions: Evidence from Three Decades of Default Cycles." Tech. rep.

Macdonald, Heidi, Jennifer Dounay-Zinth, and Sarah Pompelia. 2019. "50-State Comparison: High School Graduation Requirements." https://www.ecs.org/high-school-graduation-requirements/. Accessed: 2018-05-18, Updated: 2019-02-14.

MacKinnon, James G and Matthew D Webb. forthcoming. "Randomization inference for difference-in-differences with few treated clusters." *Journal of Econometrics* .

Maluccio, John A, John Hoddinott, Jere R Behrman, Reynaldo Martorell, Agnes R Quisumbing, and Aryeh D Stein. 2009. "The impact of improving nutrition during early childhood on education among Guatemalan adults." *The Economic Journal* 119 (537):734–763.

Marx, Benjamin M and Lesley J Turner. 2019a. "Student Loan Choice Overload." Tech. rep., National Bureau of Economic Research.

———. 2019b. "Student loan nudges: Experimental evidence on borrowing and educational attainment." *American Economic Journal: Economic Policy* 11 (2):108–41.

Mehta, Stephanie. 2016. "Meet Uber's Political Genius." *The Vanity Fair Interview. June 17, 2016* .

Molnar, Alejandro and Francis R. Ratsimbazafy. 2016. "Substituting Bikeshare for Taxis: Evidence from New York City." Working paper, Vanderbilt University.

Mundhenk, T. Nathan, Goran Konjevod, Wesam A. Sakla, and Kofi Boakye. 2016. "A Large Contextual Dataset for Classification, Detection and Counting of Cars with Deep Learning." In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science. Springer, 785–800.

Novak, Heather and Lyle McKinney. 2011. "The consequences of leaving money on the table: Examining persistence among students who do not file a FAFSA." *Journal of Student Financial Aid* 41 (3):1.

NYC. 2016. "For-Hire Vehicle Transportation Study."

NYC-DOT. 2014. "Presentation on "Protected Bicycle Lanes in NYC"." Available at nyc.gov/html/dot/downloads/pdf/2014-09-03-bicycle-path-data-analysis.pdf.

———. 2016. "Mobility Report. October 2016."

NYC-TLC. 2013. "Hail Market Analysis." Available at www.nyc.gov/html/tlc/downloads/pdf/boro_taxi_market_study.pdf.

———. 2015. "Hail Market Analysis." Available at www.nyc.gov/html/tlc/downloads/pdf/hail_market_analysis_2015.pdf.

Pinkse, Joris and Margaret E. Slade. 2010. "The Future of Spatial Econometrics." *Journal of Regional Science* 50 (1):103–117.

Poco, Jorge, Harish Doraiswamy, Huy. T. Vo, João L. D. Comba, Juliana Freire, and Cláudio. T. Silva. 2015. "Exploring Traffic Dynamics in Urban Environments Using Vector-Valued Functions." *Computer Graphics Forum* 34 (3):161–170.

Powell, Christine, Sally Grantham-McGregor, and M Elston. 1983. "An evaluation of giving the Jamaican government school meal to a class of children." *Human Nutrition. Clinical Nutrition* 37 (5):381–388.

Powell, Christine A, Susan P Walker, Susan M Chang, and Sally M Grantham-McGregor. 1998. "Nutrition and education: a randomized trial of the effects of breakfast in rural primary school children." *The American journal of clinical nutrition* 68 (4):873–879.

Razakarivony, Sebastien and Frederic Jurie. 2016. "Vehicle Detection in Aerial Imagery: A Small Target Detection Benchmark." *Journal of Visual Communication and Image Representation* 34:187–203.

Redding, Stephen J. and Matthew A. Turner. 2015. "Transportation Costs and the Spatial Organi-

zation of Economic Activity." In *Handbook of Regional and Urban Economics*, *Handbook of Regional and Urban Economics*, vol. 5, edited by Gilles Duranton, J. Vernon Henderson, and William C. Strange. Elsevier, 1339–1398.

Reilly, Vladimir, Haroon Idrees, and Mubarak Shah. 2010. "Detection and Tracking of Large Number of Targets in Wide Area Surveillance." In *Computer Vision – ECCV 2010*, Lecture Notes in Computer Science. Springer, 186–199.

Richter, Linda M, Cynthia Rose, and R Dev Griesel. 1997. "Cognitive and behavioural effects of a school breakfast." *South African Medical Journal* 87 (1 Suppl):93–100.

Rubin, Donald B. 1980. "Bias reduction using Mahalanobis-metric matching." *Biometrics* :293–298.

Sadik-Khan, Janette and Seth Solomonow. 2016. *Streetfight: Handbook for an Urban Revolution*. New York, New York: Viking, 1st ed.

Santi, Paolo, Giovanni Resta, Michael Szell, Stanislav Sobolevsky, Steven H. Strogatz, and Carlo Ratti. 2014. "Quantifying the Benefits of Vehicle Pooling with Shareability Networks." *Proceedings of the National Academy of Sciences* 111 (37):13290–13294.

Schaller, Bruce. 2017. "Unsustainable? The Growth of App-Based Ride Services and Traffic, Travel and the Future of New York City." Tech. rep., Schaller Consulting.

Schneider, Henry. 2010. "Moral Hazard in Leasing Contracts: Evidence from the New York City Taxi Industry." *Journal of Law and Economics* 53 (4):783–805.

Schwartz, Amy Ellen and Michah W Rothbart. 2017. "Let them eat lunch: The impact of universal free meals on student performance." .

Schwartz, Samuel I. 2015. *Street Smart: The Rise of Cities and the Fall of Cars*. New York, New York: PublicAffairs, 1st ed.

Scott-Clayton, Judith. 2015. "The role of financial aid in promoting college access and success: Research evidence and proposals for reform." *Journal of Student Financial Aid* 45 (3):3.

Sjoquist, David L and John V Winters. 2015. "State merit-based financial aid programs and college attainment." *Journal of Regional Science* 55 (3):364–390.

Stoddard, Christiana and Carly Urban. 2019. "The Effects of Financial Education Graduation Requirements on Postsecondary Financing Decisions." *Journal of Money, Credit, and Banking* .

Urban, Carly, Maximilian Schmeiser, J Michael Collins, and Alexandra Brown. 2018. "The effects of high school personal financial education policies on financial behavior." *Economics of Education Review* .

Wahlstrom, Kyla L and Mary S Begalle. 1999. "More Than Test Scores: Results of the Universal School Breakfast Pilot in Minnesota." *Topics in Clinical Nutrition* 15 (1):17–29.

Xie, Litian and Piotr Olszewski. 2011. "Modelling the Effects of Road Pricing on Traffic Using ERP Traffic Data." *Transportation Research Part A: Policy and Practice* 45 (6):512–522.

Zhan, Xianyuan, Samiul Hasan, Satish V. Ukkusuri, and Camille Kamga. 2013. "Urban Link Travel Time Estimation Using Large-Scale Taxi Data with Partial Information." *Transportation Research Part C: Emerging Technologies* 33:37–49.