9-2015

Essays in the Economics of Education and Health

Jesse Margolis
Graduate Center, City University of New York

How does access to this work benefit you? Let us know!
Follow this and additional works at: https://academicworks.cuny.edu/gc_etds

Part of the Economics Commons

Recommended Citation
https://academicworks.cuny.edu/gc_etds/1040

This Dissertation is brought to you by CUNY Academic Works. It has been accepted for inclusion in All Dissertations, Theses, and Capstone Projects by an authorized administrator of CUNY Academic Works. For more information, please contact deposit@gc.cuny.edu.
ESSAYS IN THE ECONOMICS OF EDUCATION AND HEALTH

by

Jesse Margolis

A dissertation submitted to the Graduate Faculty in Economics in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2015
This manuscript has been read and accepted for the Graduate Faculty in Economics to satisfy the dissertation requirement for the degree of Doctor of Philosophy.

Michael Grossman

Date

Chair of Examining Committee

Wim Vijverberg

Date

Executive Officer

Supervisory Committee
Michael Grossman
David Jaeger
Theodore Joyce

THE CITY UNIVERSITY OF NEW YORK
Abstract

ESSAYS IN THE ECONOMICS OF EDUCATION AND HEALTH

by

Jesse Margolis

Advisor: Professor Michael Grossman

In this work, I study several aspects of the economics of education and health. The first chapter focuses on education, analyzing the New York City Teacher Bonus Program. Recent reviews of the program have found lower student test scores at schools randomly chosen to receive performance pay than at control schools not eligible for bonuses. Several studies have suggested lower intrinsic motivation among teachers at treatment schools as a possible mechanism driving these surprising results. The program was abruptly ended after three years, allowing me to use post-trial data to study the program’s persistent effect on intrinsic motivation absent any effect on extrinsic motivation. While I replicate prior results showing that the teacher bonus program had a negative impact on student test scores, a difference-in-difference analysis of teacher survey responses indicates that this was not likely caused by a change in teachers’ intrinsic motivation. Moreover, a regression discontinuity (RD) study demonstrates that the observed “negative” effect on test scores was not driven by a decline in performance at treatment schools, but rather by an increase in performance at control schools. This finding highlights a risk with RCT experiments in the social sciences: even with proper randomization, the control group may not be a valid counterfactual for the treatment group.

The second chapter – written with coauthors Jason Hockenberry, Michael Grossman, and Shin-Yi Chou – focuses on health. We analyze patients’ behavioral response to more
invasive surgery. Over the last several decades, numerous medical studies have compared the effectiveness of two common procedures for Coronary Artery Disease: Percutaneous Coronary Intervention (PCI) and Coronary Artery Bypass Graft (CABG). Most evidence indicates that CABG – the more invasive procedure – leads to superior long term outcomes for otherwise similar patients, though there is little consensus as to why. In this article, we propose a novel explanation: patient offsetting behavior. We hypothesize that patients who undergo the more invasive procedure, CABG, are more likely to improve their behavior – eating, exercise, smoking, and drinking – in a way that increases longevity. To test our hypothesis, we use Medicare records linked to the National Health Interview Survey to study one such behavior: smoking. We find that CABG patients are 12 percentage points more likely to quit smoking in the one-year period immediately surrounding their procedure than PCI patients, a result that is robust to numerous alternative specifications.

The third chapter combines education and health, studying the effect of schools on students’ weight. In this chapter, I use New York City’s high school admissions process as a natural experiment to assess the impact of going to a particular school on student weight, BMI, and probability of being overweight or obese. For each high school that admitted students by an oversubscribed lottery, I compare the weight (or BMI, overweight, or obesity) of students admitted to a particular school with the weight of students who applied, but were not admitted based purely on a random number. Analyzing over 100 lotteries, I find statistically significant school effects on student weight about as often as one would expect by chance. I conclude that within NYC the specific high school a student attends is unlikely to have a substantial effect on his or her weight.
Acknowledgements

This work would not have been possible without the help and guidance of many people, starting with my dissertation committee. I would like to thank my advisor, Michael Grossman, who has supported me in many ways over the past five years. Mike has taught me about health economics, given me excellent research advice, and – not surprisingly to anyone who knows him – helped me get a job. Not only is Mike a renowned economist, but he is a patient mentor and a passionate advocate for all of his students. I would also like to thank my co-advisor David Jaeger, a brilliant applied econometrician whose weekly research brown-bag – at which I presented each chapter of my dissertation – provided me some of the most helpful feedback I have received. The third member of my committee, Ted Joyce, is an inspiring teacher and researcher who gave me good advice and sets an example – as both an educator and scholar – that I hope to emulate.

There are many others at the CUNY Graduate Center (GC) to whom I am grateful. Merih Uctum and Wim Vijverberg were kind enough to admit me to the GC’s Economics Ph.D. program, which they have ably administered and worked tirelessly to improve. I learned so much from the faculty and appreciated the hard work of the GC staff, particularly Diana Santiago, who kept the program running smoothly and was always available and happy to answer any student’s question. My classmates over the past five years made my doctoral experience truly rewarding. We learned together, we worked together, and we made friendships that will last a lifetime. I am lucky to have been part of such a fantastic cohort.

Finally, I would like to thank my family. My parents, Jim and Angela, my siblings, Carrie, Gabe, Emily, Juli, Bryan, and Steve, and my in-laws, John and Connie, have all encouraged and supported my mid-career shift into Academia. My boys, James and Luke, have
tolerated their dad working too many late nights and have enjoyed visits to the GC and Brooklyn College. And last, but certainly not least, thank you Shannon. You have been there for me from the beginning to the end. From reading my application essays to suggesting improvements to my dissertation, I have come to rely on your sound advice, good judgment, and unfailing support. This work is dedicated to you.
Contents

1. Does Performance Pay Reduce Teachers’ Intrinsic Motivation? Evidence from the New York City Teacher Bonus Program ........................................................................................................ 1
   Jesse Margolis

1.1 Introduction ....................................................................................................................... 2
1.2 Methodology & Data ......................................................................................................... 8
   1.2.1 Student Test Scores (Randomized Controlled Trial) ............................................... 10
   1.2.2 Teacher Survey Responses (Difference-In-Differences) .......................................... 11
   1.2.3 Regression Discontinuity Design ............................................................................. 13
1.3 Descriptive Results ....................................................................................................... 15
   1.3.1 Confirmation of Randomization ............................................................................. 15
   1.3.2 Student Test Scores ............................................................................................... 17
   1.3.3 Teacher Surveys .................................................................................................... 18
   1.3.4 Regression Discontinuity Design ........................................................................... 20
1.4 Results ............................................................................................................................ 22
   1.4.1. Student Test Scores ............................................................................................ 22
   1.4.2 Teacher Survey Responses ..................................................................................... 25
1.5 Conclusion ....................................................................................................................... 34

2. Moral Hazard and Less Invasive Medical Treatment for Coronary Artery Disease: The Case of Cigarette Smoking ............................................................................................. 37
   Jesse Margolis, Jason Hockenberry, Michael Grossman, and Shin-Yi Chou

2.1 Introduction ....................................................................................................................... 38
2.2 Data .................................................................................................................................. 42
2.3 Initial Analysis .................................................................................................................. 44
2.4 Results ............................................................................................................................. 50
   2.4.1 Discrete-time Linear Probability Hazard Model (Individual Data) ......................... 52
   2.4.2 Multi-Period Quit Model Using Grouped Data ..................................................... 56
2.5 Discussion ......................................................................................................................... 60
   2.5.1 Selection Bias ......................................................................................................... 60
   2.5.2 Alternative Mechanisms ......................................................................................... 61
   2.5.3 Quitting and the PCI-CABG Mortality Differential .............................................. 64
2.6 Conclusion ........................................................................................................................ 66
3. Schools and Obesity: A Natural Experiment Using the New York City High School Admissions Process ................................................................................................................. 67

Jesse Margolis

3.1 Introduction .................................................................................................................. 68
3.2 Methodology ............................................................................................................... 70
3.3 Data ............................................................................................................................. 72
3.4 Confirming Randomization in School Lotteries ......................................................... 73
3.5 Results ........................................................................................................................ 78
3.6 Conclusion and Discussion ......................................................................................... 84

4. Appendix ....................................................................................................................... 87

5. Bibliography .................................................................................................................. 95
List of Tables

Chapter 1
Table 1.1 – Expected Effect of the Bonus Program on Test Scores and Survey Results ............. 9
Table 1.2 – Pre-Bonus Program Characteristics at Treatment and Control Schools .................. 16
Table 1.3 – Teacher Response Rates to the NYC School Survey ........................................... 19
Table 1.4 – Impact of the Bonus Program on Average Student Test Scores ......................... 23
Table 1.5 – Impact of the Bonus Program on Teacher Agreement with the Statement: “I usually look forward to each working day at my school” .......................................................... 25
Table 1.6 – Intrinsic Motivation Questions vs. Other Questions (Subjective Division) .......... 26
Table 1.7 – Intrinsic Motivation Questions vs. Other Questions (Factor Analysis Division) ..... 28
Table 1.8 – Impact of Bonus Eligibility on Student Test Scores (Regression Discontinuity) .... 32
Table 1.9 – Comparison of RD and RCT Results ................................................................. 33

Chapter 2
Table 2.1 – Patient Characteristics by Treatment ................................................................... 45
Table 2.2 – Smoking Status as of NHIS Interview Date .......................................................... 46
Table 2.3 – Quitting Within Six Months of Treatment ............................................................. 47
Table 2.4 – Regression Results for Discrete Time Linear Probability Hazard Model ............ 55
Table 2.5 – Quit Rate Regression with Grouped Data ............................................................ 59

Chapter 3
Table 3.1 – FITNESSGRAM Participation by Year ................................................................. 73
Table 3.2 – Number of Programs with First-Choice Lotteries by Year ..................................... 75
Table 3.3 – Covariate Balance for High School A ................................................................. 76
Table 3.4 – Fraction of Lotteries with Covariate P-Values Less than or Equal to 0.05 ............ 78
Table 3.5 – Outcomes for High School A ............................................................................. 79
Table 3.6 – Fraction of Lotteries with Outcome P-Values Less than or Equal to 0.05 .......... 82

Appendix
Table A.1.1 – Subjective Division of Teacher Survey Questions ........................................ 87
Table A.1.2 - Intrinsic Motivation Questions vs. Other Questions (Subjective Division, Controls for 2007) ................................................................. 89
Table A.1.3 - Intrinsic Motivation Questions vs. Other Questions (Factor Analysis Division, Controls for 2007) ........................................................................................................................................... 90
Table A.1.4 – Teacher Survey Triple Difference Regression ................................................................................................................................. 91
Table A.1.5 – Impact of Bonus Eligibility on Student Test Scores (RD, Local Linear Regression) ........................................................................................................................................... 92
Table A.2.1 – NHIS/Medicare Data Linkage ......................................................................................................................................................... 93
Table A.2.2 – Regression Results for Discrete Time Linear Probability Hazard Model, Testing for Significance of Pre-Treatment Differences in Quit Rate ................................................................................. 94
List of Figures

Chapter 1
Figure 1.1 – Test Scores Distributions Before (2007) and After (2011) the Bonus Program ..... 18
Figure 1.2 – Probability of Treatment by Distance from Peer Index Cut Point .......................... 20
Figure 1.3 – Test Scores by Distance from Peer Index Cut Point ........................................... 21
Figure 1.4 – Test Scores by Distance from Peer Index Cut Point (2011) ................................. 30

Chapter 2
Figure 2.1 – Comparative Effectiveness: PCI vs. CABG (Weintraub et al., 2012) ....................... 40
Figure 2.2 – Smoking Rate by Year Relative to Diagnosis (MM) or Procedure (PCI & CABG) 48
Figure 2.3 – Quit Rate by Year Relative to Diagnosis (MM) or Procedure (PCI & CABG) ...... 49

Chapter 3
Figure 3.1 – Distribution of PSAL Sports teams by school and incoming obesity rates by program in New York City ........................................................................................................... 69
Figure 3.2 – P-Values for Average Weight in 8th Grade ............................................................. 77
Figure 3.3 – P-Values for Average Weight Three Years After Starting High School ............. 81
Figure 3.4 - Fraction of Lotteries with Outcome P-Values Less than or equal to 0.05 .......... 83
1. Does Performance Pay Reduce Teachers’ Intrinsic Motivation? Evidence from the New York City Teacher Bonus Program

Jesse Margolis
1.1 Introduction

Merit pay, or pay-for-performance, is common in many industries (Lemeiux, MacLeod, and Parent, 2009). Salespeople are often paid on commission, waiters receive a substantial portion of their pay in tips, and bankers receive year-end bonuses based, in part, on their supervisors’ assessment of their performance. By contrast, public school teachers in the U.S. are usually paid according to a uniform, district-wide pay scale that takes into account years of experience, level of education, and little else. Over the last decade, many schools and districts have started to experiment with merit pay, often paying teachers bonuses based on their students’ test scores. In 2009, the U.S. Department of Education said it would judge schools applying for the Federal Race to the Top grants, in part, on the degree to which they used teacher and principal evaluations “to provide opportunities for highly effective teachers and principals...to obtain additional compensation...”\(^2\) When the Round 1 winners – Tennessee and Delaware – were announced on March 29, 2010, both states had teacher bonus programs as part of their proposals.

Internationally, evaluations of teacher bonus programs have found largely positive impacts on student test scores and other outcomes (see Lavy, 2009 in Israel; Glewwe, Ilias, and Kremer, 2010 in Kenya; and Muralidharan and Sundararaman, 2011, in India). However, within the United States, the two large teacher bonus programs that have been rigorously evaluated have shown no positive impact on student test scores or other outcomes. In Nashville, the Project on Incentives in Teaching (POINT) program provided middle school math teachers with the potential to win individual bonuses of up to $15,000 annually based on improvement in their

---

1 I would like to thank Michael Grossman, David Jaeger, Ted Joyce, Steve O’Connell, seminar participants at the CUNY Graduate Center, and attendees at the Association for Education Finance and Policy Annual Conference for helpful comments on an earlier version of this research.
2 Federal Register, November 18, 2009, pp. 59836-59872
students’ math test scores. After three years, students in randomly selected treatment classrooms – whose teachers were eligible for bonuses – showed no improvement in test scores over students in control classrooms (Springer et al., 2010). In New York City, the School-Wide Performance Bonus program – hereafter referred to as the NYC teacher bonus program – provided teachers and other educators at high-needs schools with the opportunity to win a roughly $3,000 annual bonus based on their school’s performance on the Progress Report (an annual evaluation of test scores and other outcomes). After three years, students at schools randomly selected to be in the treatment group showed no better – and in some cases, worse – outcomes on standardized tests and other measures (Marsh et. al., 2011; Fryer, 2013; Goodman and Turner, 2013).

These results run counter to classical economic theory, which posits a strong link between performance-based-pay and improved performance. In traditional economic models, both greater consumption and greater leisure are assumed to increase utility. If one’s consumption does not increase with greater work effort – since pay is unrelated to effort – then one will work as little as possible to maximize leisure. Researchers in social psychology, however, have theorized that people are intrinsically motivated to undertake many activities and providing performance-based pay for an intrinsically motivating activity can be counterproductive. Performance-based pay can decrease intrinsic motivation and potentially performance, a finding that has been repeatedly demonstrated in laboratory settings.

In a pioneering study, college students were provided with the opportunity to work on a series of challenging puzzles, and half of the students were paid based on the number of puzzles they solved. Those who were paid proved to be less motivated to work on these puzzles at a later point, during their free time, than students who were unpaid (Deci, 1971). Edward Deci, who
conducted the study, said it “supported the hypothesis that if monetary rewards are given to subjects for doing an intrinsically motivated activity, and if the rewards are made contingent on their performance, their intrinsic motivation for the activity will decrease” (Deci, 1975, p. 132). In a related study, Uri Gneezy and Aldo Rustichini find that compared to an unpaid control group, college students do worse on an IQ test if paid a small sum of money per right answer, though better if paid a large sum (2000a).

One field study showing behavior consistent with a loss of intrinsic motivation is sometimes called the Israeli Daycare Study (Gneezy and Rustichini, 2000b). In an effort to encourage parents to pick up their children on time, ten daycare centers in Israel began levying a small fine on those who were late. After levying the fine, the proportion of parents arriving late to pick up their children increased, prompting the centers to remove the fine. However, even when the fine was removed, the proportion of parents arriving late stayed at the new higher level.

The design of the New York City teacher bonus program offers an opportunity to test for this phenomenon – financial incentives reducing intrinsic motivation – in a rigorous, real-world setting. Starting in the 2007/08 school year – hereafter referred to as 2008 – the NYC teacher bonus program was run as a three-year randomized controlled trial, with 212 treatment schools and 184 control schools (of which 175 treatment schools and 127 control schools – the focus of this study – had elementary or middle school testing data in all years between 2006 and 2013). The entire group of 396 schools was not, itself, a random sub-sample New York City’s 1,200+ schools. Rather, it included the 396 highest-need schools, as defined by their Progress Report peer index, a continuous measure based on student demographic characteristics and prior test scores.³

---

³ The peer index is intended to be a measure of fixed student characteristics at a school. At elementary and K-8 schools, the peer index is based on the percentage of students who require special education services, are Black or
At treatment schools, all educators affiliated with the local teachers’ union (the United Federation of Teachers) were eligible to receive bonuses averaging $3,000 per person if their school met its target on the annual New York City Progress Report. At the K-8 level, a school’s score on the Progress Report was determined largely by students’ test scores and growth in test scores (85%), attendance (5%) and the results of parent, teacher, and student surveys (10%). At the high school level, a portion of the weight was removed from test scores and placed on a school’s graduation rate and students’ credit accumulation.

Educators at 62% of treatment schools won a bonus during the first year of the program (2008), a rate that rose to 88% during the second year (2009), and 13% during the third year (2010). Over the three years of the program, New York City paid out nearly $56 million in bonuses (Marsh et al. 2011). On January 20, 2011, the New York City Department of Education (NYCDOE) announced it was suspending the teacher bonus program, based on uncertain benefits and budget constraints. In July, 2011, after the RAND Corporation released its final report, the NYCDOE permanently ended the teacher bonus program (Marsh et al., 2011). After the conclusion of the three-year pilot, 2011 was the first school year in which teachers at treatment schools no longer had the possibility of receiving a performance bonus.

In explaining the neutral to negative results of the New York City bonus program, researchers have pointed to a decrease in intrinsic motivation as one possible cause. Referring to their finding of a negative and statistically significant effect of the bonus program on math test scores at large schools, Goodman and Turner (2013) say that “One explanation is that the bonus...”

Hispanic, are English Language Learners, or qualify for free or reduced price lunch. At the middle school level, the peer index is based on the average Math and ELA test scores students earned in 4th grade, prior to entering middle school. At the high school level, the peer index is based on the average Math and ELA test scores students earned in 8th grade, prior to entering high school.

4 A committee of two teachers and two principal appointees determined the exact distribution of the bonus among all eligible recipients. In 2008, 52% of staff who received an award received exactly $3,000 (Marsh et. al., 2011)
program crowded out teachers’ intrinsic motivation…” Fryer (2013), who finds a negative effect on the bonus program on middle school math and language arts test scores, says that “…some argue that teacher incentives can decrease a teacher’s intrinsic motivation…” And Marsh et al (2011), who conducted extensive interviews with New York City teachers as part of the program’s official evaluation, report that “In many schools, staff members also typically attributed their hard work and their efforts to improve their practices to intrinsic motivations far above any external pressures or incentives.”

In my study, I seek to answer the following question: did the New York City teacher bonus program lower teachers’ intrinsic motivation? I do this by extending prior research in three ways. First, I use the RCT assignment to assess the impact of the bonus program on student test scores in the three years following the suspension of the program. To the extent the bonus has a persistent negative effect on intrinsic motivation – as prior literature indicates it might – studying the post-suspension period allows me to disentangle this effect from any positive effect on extrinsic motivation during the program period itself. Second, I conduct an item-level analysis of the NYC teacher survey using all questions asked consistently before, during, and after the teacher bonus program. Using a difference-in-differences (DD) methodology, I assess the impact of the bonus program on teacher responses to questions related to intrinsic motivation and compare this to the impact of the bonus program on questions unrelated to intrinsic motivation. Finally, I complement the RCT with a regression discontinuity (RD) study of student test scores, taking advantage of the fact that the entire experimental sample – both treatment and control schools – represent the highest-need one-third

---

7 The canonical DD study has one first difference over time, whereas the two differences I note here are between the treatment and control group and between questions related and unrelated to intrinsic motivation. I address the time dimension in my study in two ways. First, I estimate a model where I control for the difference in survey scores in 2007, before the bonus program was introduced. Second, I estimate a triple difference model, where the third difference is over time. The results, which are shown in an appendix, are similar.
of schools in New York, as defined by a continuous index with a rigid cut point. This allows me to use non-eligible schools as a counterfactual for both the treatment and control schools and answer the following question: do we observe a negative effect of the bonus program because the treatment schools went down relative to their (unobserved) counterfactual or, rather, because the control schools improved?

In analyzing the post-program period, I find that the negative effects of the bonus program on student test scores continued and grew after the program was suspended. The bonus program caused treatment schools to perform between 0.13 and 0.17 standard deviations worse than control schools in math and between 0.08 and 0.13 standard deviations worse in language arts in the three years following the program’s suspension. While this result, taken alone, is consistent with a decrease in intrinsic motivation among teachers at treatment schools, the next two analyses provide evidence against such an interpretation. In analyzing the teacher survey, I again show that the bonus program had an effect: teacher survey scores at treatment schools were lower during the post-program period. However, the trend is the same for questions related and unrelated to intrinsic motivation, a pattern one wouldn’t expect if changes in intrinsic motivation were driving this result. Finally, in the RD analysis – where I use non-eligible schools as a counterfactual for both the treatment and control schools – I replicate the main results from the RCT analysis: students at treatment schools had lower test scores than students at control schools, especially in the period following the suspension of the bonus program. However, this does not appear to be caused by a decline in performance at treatment schools, but rather by an improvement at control schools. This finding is inconsistent with a decline in intrinsic motivation among teachers at treatment schools and suggests that another mechanism may be

---

8 These effects are relative to the standard deviation of the school-level distribution of mean test scores in New York City.
driving the NYC teacher bonus program results observed in this and other studies. Overall, I find little evidence to suggest that intrinsic motivation declined at schools participating in the New York City teacher bonus program.

1.2 Methodology & Data

In this paper, I analyze the impact of the NYC teacher bonus program on teachers’ intrinsic motivation in three ways. First, I take advantage of the fact that the program was designed as a Randomized Controlled Trial (RCT) to assess its impact on student test scores. I compare average school-wide student test scores in math and language arts in treatment schools to those in control schools. Second, I use a difference-in-differences methodology to study the impact of the program on teacher attitudes. Making use of the RCT, I assess the impact of the bonus program on teacher responses to survey questions related to intrinsic motivation and compare that to the impact of the bonus program on teacher survey questions unrelated to intrinsic motivation. Third, I use a Regression Discontinuity (RD) design to better understand the mechanism behind the observed RCT results. In particular, I analyze whether the test score results observed in the RCT are due to a decline in performance at treatment schools or an improvement in performance at control schools.

In each analysis, I study the impact of the bonus program over eight years, which are divided into three distinct periods in Table 1.1. In the pre-program period, comprised of the years 2006 and 2007, I expect no impact of the bonus program on test scores or teacher attitudes because the program had not yet been announced. Analyzing these years serves to validate the randomization and provide a falsification test. During the three-year program period (2008 to 2010), even if the program reduced teachers’ intrinsic motivation, the predicted impact on both
student test scores and teacher survey results is indeterminate. For student test scores, any negative impact of the program due to teachers’ lower intrinsic motivation might have been offset by a positive impact of the reward on teachers’ extrinsic motivation. For teacher survey results, the tendency to report lower levels of intrinsic motivation might have been offset by the desire to actually win a bonus; teacher survey results made up 3-5% of a school’s Progress Report score, on which the bonus was based. Once the bonus program was suspended in 2011, the extrinsic reward was removed, leaving only any persistent change in intrinsic motivation. As indicated by the results of the Israeli Daycare Study, once a person’s intrinsic motivation has been lowered by an external reward/penalty, it may not quickly return to its previous higher level. If true in the case of the New York City teacher bonus program, we would be most likely to observe results consistent with a decrease in teachers’ intrinsic motivation after the bonus was removed.

Table 1.1 – Expected Effect of the Bonus Program on Test Scores and Survey Results

<table>
<thead>
<tr>
<th>Period</th>
<th>Years</th>
<th>Expected Effect</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-program</td>
<td>2006</td>
<td>None</td>
<td>Program not yet announced</td>
</tr>
<tr>
<td></td>
<td>2007</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Program</td>
<td>2008</td>
<td>Indeterminate</td>
<td>Positive effect of extrinsic reward countered by negative effect</td>
</tr>
<tr>
<td></td>
<td>2009</td>
<td></td>
<td>of reduced intrinsic motivation</td>
</tr>
<tr>
<td></td>
<td>2010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-program</td>
<td>2011</td>
<td>Negative</td>
<td>Loss of intrinsic motivation only (to the extent it persists)</td>
</tr>
<tr>
<td></td>
<td>2012</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2013</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: each year listed refers to the end of the school year (i.e. 2013 represents the 2012/13 school year)
1.2.1 Student Test Scores (Randomized Controlled Trial)

In line with prior research on the New York City teacher bonus program, I focus first on student test scores in math and ELA in grades 3 through 8. In addition to connecting with prior literature, student test scores are a useful starting point for two reasons. First, a large body of research has established that teachers have a strong influence on student test scores, so it is plausible to think that a change in teachers’ intrinsic motivation would be reflected in their students’ scores (Rockoff, 2004, Rivkin, Hanushek, and Kain, 2005, Kane and Staiger, 2008). Second, for many policy makers, a change in teachers’ intrinsic motivation is important only if it affects student outcomes (of which test scores are one measure). Assuming policy makers are primarily concerned with improving student outcomes, and only secondarily with the welfare of the education workforce, a change in teacher attitudes that affects student performance will be more important than one that does not.

In New York State, all third through eighth graders who are not severely disabled take standardized tests in the second half of the year. For the years 2006 through 2013, the NYCDOE web site provides information on average test scores and number of students tested by grade by school. I create a single average test score variable for each school in each year, weighting by the number of students in each grade. To create the average test score variable, I normalize the scaled scores within each grade to have mean zero and standard deviation one – based on the citywide distribution of average school-level test scores – before combining into a single school-level average. In the results below, I restrict my sample to those schools that have ELA test score data for all years in the study period: 2006 to 2013. This excludes two schools that opened during this period and ten schools that closed during this period. My results are robust to including these schools.
To assess the impact of the NYC teacher bonus program on student test scores, I fit the simple linear model shown in Equation 1.1 using Ordinary Least Squares (OLS).

\[ Test_{st} = \beta_0 + \beta_1 D_s + \beta_2 Test_{s,t=2007} + \beta_3 X_{st} + \epsilon_{st} \] (1.1)

In Equation 1.1, \( Test_{st} \) represents the average school-wide test score at school \( s \) in year \( t \), \( D_s \) is a dummy variable equal to one if school \( s \) was invited to be in the treatment group and zero if school \( s \) was placed in the control group, \( Test_{s,t=2007} \) is the pre-treatment average test score, and \( X_{st} \) represents a vector of school-level control variables. Assuming random assignment to the treatment group, \( \beta_1 \) represents the causal effect on student test scores of being invited to participate in the teacher bonus program. Since not all schools invited to participate chose to do so, \( \beta_1 \) should be viewed as an Intention to Treat (ITT) estimate.\(^9\) Neither \( Test_{s,t=2007} \) nor \( X_{st} \) are necessary to identify the causal impact of the bonus program on student test scores. They are included only to reduce residual variance and allow for a more precise impact estimate. In the base specification, I restrict both \( \beta_2 \) and \( \beta_3 \) to be zero.

1.2.2 Teacher Survey Responses (Difference-In-Differences)

To assess the impact of the teacher bonus program on a more direct measure of intrinsic motivation, I use teacher responses to the New York City School Survey. Every year, in March and April, all parents, all teachers, and students in grade 6 and above are invited to fill out the NYC School Survey. When the surveys were first collected in April, 2007, the response rate was 26% for parents, 44% for teachers, and 65% for students. The surveys collected in April, 2013,  

\(^9\) To be part of the program, 55% of UFT represented staff had to vote to participate. According to Marsh et. al. (2011), 87% of schools invited to be part of the bonus program elected to do so.
had a response rate of 54% for parents and 83% for teachers and students. On the teacher survey, many questions have been asked consistently for most or all of the time period under study. From 2007 to 2012, for example, teachers were asked for their agreement with the statement: “Teachers in my school recognize and respect colleagues who are the most effective teachers.” Teachers could respond that they Strongly Agreed, Agreed, Disagreed, or Strongly Disagreed with this statement. For each question in each school, the NYCDOE assigned a score from 0 to 10 along a Likert-like scale, with 0 corresponding to Strongly Disagree and 10 corresponding to Strongly Agree (or the reverse, when Strongly Agree was a negative response).

Collections of questions were grouped into four categories – academic expectations, communication, engagement, and safety – and average scores for each category were calculated. For each category, the average score across all respondent groups – teachers, students, and parents – appears on the Progress Report and counts for 10% of a school’s letter grade. Detailed item-level data by school are made available on the NYCDOE’s web site.

I focus on the 45 questions that were asked consistently between 2007 and 2012. I exclude 2013 since the survey was redesigned that year and the majority of questions changed. I identify questions related to intrinsic motivation in two ways. First, I review the questions and subjectively divide them into nine questions that I consider to be most closely related to intrinsic motivation and 36 remaining questions (see Table A.1.1 in the appendix). Second, I select the one question on the 2012 survey that was most closely related to the items on widely-used *Intrinsic Motivation Inventory* (Deci & Ryan, 2014): “I usually look forward to each working day at my school.” Since this question was only asked in 2012 and 2013, I develop a broader group of intrinsic motivation-related questions through an exploratory factor analysis of survey results in 2012. Following Deci & Ryan (2014), I identify three categories of questions by
grouping together items that have a factor loading of at least 0.6 on one category and less than 0.4 on the others. I consider the category that contains the “look forward” question as the one most closely related to intrinsic motivation. The other two categories appear to contain questions that are broadly related to teacher collaboration and school safety. To estimate the effect of the bonus program on teachers’ survey responses, I fit the equation shown below using Ordinary Least Squares:

\[
z_{st} = \beta_0 + \beta_1 D_s + \beta_2 z_{s,t=2007} + \epsilon_{st} \tag{1.2}
\]

where

\[
z_{st} = x_{st} - y_{st} \tag{1.3}
\]

Here \(x_{st}\) is the average teacher score on the intrinsic motivation questions for school \(s\) in year \(t\), \(y_{st}\) is the average score on non-intrinsic motivation questions, and \(z_{st}\) is the difference between the two. With \(\beta_2\) set to zero, \(\beta_1\) provides the straightforward difference-in-differences (DD) estimate where the first difference is between the two types of survey questions and the second difference is between the treatment and control group. Unlike the canonical DD model, neither difference is related to time, though I add a time dimension in two ways. First, I include the pre-treatment value of the difference in survey scores, \(z_{s,t=2007}\), as a variable on the right-hand side of the equation. Second, I recalculate the dependent variable as the difference between the current year difference and the 2007 difference, thus estimating a triple-difference model.

1.2.3 Regression Discontinuity Design

To validate and extend the RCT results, I analyze the student test score data using a
Regression Discontinuity (RD) Design, taking advantage of the rigid discontinuity in school eligibility for the bonus program. Prior to randomization, 430 high-need schools were selected to be eligible for the bonus program based on their peer index on the NYC Progress Report.\footnote{Note that my study focuses on the 302 elementary, middle, and K-8 schools that had consistent testing data from 2007 to 2013 and were not barred from participation by the UFT (discussed below).} These bonus-eligible schools – a subset of all 1,217 NYC schools that received a Progress Report in 2007 – were selected entirely based on their Progress Report peer index. Within each of five school types – elementary, K-8, middle, high, and transfer – schools above a particular peer index score were chosen to be eligible for the teacher bonus program and schools below that score were ineligible.\footnote{For middle schools, high schools, and transfer schools, a lower peer index meant a higher-need school. Technically, therefore, schools below a certain peer index threshold were eligible for the bonus program. In this analysis, I normalize all peer indices by calculating z-scores within school type and then defining a higher z-score to be a higher-need schools (multiplying by -1 where needed).} Conceptually, an RD design compares bonus-eligible schools just above the peer index cut point to bonus-ineligible schools just below the cut point. Assuming that bonus-ineligible schools just below the cut point present a valid counterfactual for bonus-eligible schools just above the cut point, once the peer index (i.e. forcing variable) is controlled for, any difference in outcomes reflects the impact of being eligible for the bonus program.

I use several specifications to implement the RD design based on the equation below:

\[
Test_{st} = \beta_0 + \beta_1 E_s + \beta_2 f(PeerIndex_{s,t=2007}) + \beta_3 E_s \times f(PeerIndex_{s,t=2007}) + \epsilon_{st} \tag{1.4}
\]

In Equation 1.4, \(Test_{st}\) is the average math or ELA test score for school \(s\) in year \(t\), \(E_s\) is an indicator for whether school \(s\) was among original 430 bonus-eligible schools with one indicating eligible and zero indicating ineligible, and \((PeerIndex_{s,t=2007})\) is a polynomial function of the school’s peer index in 2007, the forcing variable by which school eligibility for the bonus program was determined. The polynomial function is interacted with \(E_s\) to allow it to
be fit separately on either side of the eligibility cut point.

In keeping with recommendations in the RD literature, I estimate various versions of this equation (Lee and Lemieux, 2010). I allow $f(PeerIndex_{s,t=2007})$ to vary from a first through fourth order polynomial function and I restrict the sample to bandwidths increasingly close to the cut point. I also run a local linear regression using a triangular kernel (Nichols, 2011). In the tables presented in the main paper, I focus on the results using all of the data where $f(PeerIndex_{s,t=2007})$ is a linear function. I do this for two reasons. First, as shown in Figure 1.3, there appears to be a strong linear relationship between peer index and test scores. Second, the linear regression performs best on “placebo” tests using data from 2006 and 2007, prior to the implementation of the bonus program. Results from the local linear regression – which are very similar to those from a full sample linear regression and perform only slightly worse on placebo tests – are presented in an appendix.

1.3 Descriptive Results

1.3.1 Confirmation of Randomization

Table 1.2 shows basic demographic and performance characteristics for the treatment and control schools during the period prior to the bonus program. The schools in the study have a high proportion of Black or Hispanic students (96%) and a large share of students who qualify for free or reduced price lunch. The average test scores for these schools are substantially below the New York City mean, with z-scores of between -0.75 and -0.78, depending on the subject and the group. This lower-than-average performance in the pre-treatment period is consistent with the eligibility requirement for the bonus program: schools had to have among the highest peer index scores (i.e. highest need) on the 2007 Progress Report.
When comparing the 175 treatment schools to the 127 control schools, we see that the two groups are balanced along each dimension. The p-value of a test for the equality of means is far from traditional levels of statistical significance for each variable, giving confidence in the randomization. This aligns with prior research into the New York City teacher bonus program, which finds evidence consistent with proper randomization (Fryer, 2013; Goodman and Turner, 2013; Marsh et. al., 2011).

12 The randomization of schools into the treatment and control groups was conducted by Roland Fryer, who at the time was serving as the Chief Equality Officer of the NYCDOE.
1.3.2 Student Test Scores

Figure 1.1 shows the distribution of school-wide average math and ELA test scores in the treatment and control groups. As before, the test scores are measured in z-scores, standardized based on the NYC-wide mean and school-level standard deviation within each year/grade/subject combination. The left side of the chart shows the distribution in 2007, prior to the implementation of the bonus program. In both subjects, the treatment and control groups appear to be fairly similar to one another. Both distributions are centered below zero, consistent with the selection criteria for the program.

The right side of the figure shows the distributions in 2011, the first year after the bonus program was suspended. We can see that both the treatment and control distributions have widened when compared to 2007. In particular, the distributions have expanded to the right somewhat, with a larger number of schools having above average test scores (z-score > 0). Given that schools were selected to be eligible for the study based on their low incoming achievement levels and high-need demographic characteristics from 2007, it makes sense that by 2011, the schools would be less similar to one another (within each group). Since both groups of schools were, by definition, among the lowest achieving schools in 2007, their standing relative to all NYC schools would tend to improve over time due to simple mean reversion.

We can also see that by 2011, the treatment group distribution is below the control group distribution. This is true for both subjects and appears to be true across the majority of the distribution (i.e. the shape of the two graphs are fairly similar). While clearly visible, the difference between the two distributions is small, representing what appears to be a fraction of standard deviation in test scores.
1.3.3 Teacher Surveys

In addition to validating randomization, when analyzing the teacher survey, we must also address the possibility of response bias. While it is administered as a census, teachers are not required to take the survey and not all do. Since it was first administered in 2007, the overall teacher response rate has increased from 44% to 83%. To test whether the bonus program itself had an influence on teacher response rates, I run a simple regression along the lines of Equation 1.2, where $z_{st}$ represents each school $s$’s response rate in year $t$. The results are shown in Table 1.3.
Table 1.3 – Teacher Response Rates to the NYC School Survey

<table>
<thead>
<tr>
<th></th>
<th>Pre-Pgm.</th>
<th>Program</th>
<th>Post-Program</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>$D_i$ Coefficient</td>
<td>2007</td>
<td>2008</td>
<td>2009</td>
</tr>
<tr>
<td>Teacher, No Controls</td>
<td>-0.029</td>
<td>-0.014</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td>(0.026)</td>
<td>(0.025)</td>
</tr>
<tr>
<td>Teacher, Control for 2007</td>
<td>-0.002</td>
<td>0.036</td>
<td>0.047**</td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
<td>(0.024)</td>
<td>(0.024)</td>
</tr>
</tbody>
</table>

N | 299 | 302 | 302 | 302 | 302 | 299 | 302

Note: each coefficient is the result of a separate regression with school-level teacher survey response rate as the dependent variable and treatment ($D_i$) as the independent variable. Regressions with controls adjust for the survey response rate in 2007. Regressions are weighted by the number of responders at each school. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

In Table 1.3, the coefficients show the percentage point difference in response rates between the treatment and control groups. When looking at the regression without controls in the first row, none of the differences are statistically significant. We can see, however, that the point estimate for the treatment effect in 2007 – prior to the start of the bonus program – is negative, while the point estimate for 2010 is positive. In the second row, I control for 2007 survey scores in each regression, which leads the 2010 coefficient to become a positive and statistically significant 4.7 percentage points. Overall, there is some evidence that teacher response rates increased at bonus-eligible schools during the bonus program – perhaps unsurprising since teacher surveys played a small role in determining whether a school won a bonus – but no evidence of a lasting effect. In the three years since the bonus program ended, the difference between teacher response rates at treatment and control schools has been indistinguishable from zero.
1.3.4 Regression Discontinuity Design

A necessary condition for the validity of an RD design is that the forcing variable influences assignment to treatment. Figure 1.2 shows this to be the case for the NYC teacher bonus program. The x-axis shows the distance from the peer index cut point in units of standard deviation, where schools are grouped into bins of width 0.1. The y-axis shows the percentage of schools in each bin that were invited to participate in the bonus program. No schools below the eligibility cut point were invited to participate in the bonus program and slightly more than 50% of schools above the eligibility cut point were invited to participate.

Figure 1.2 – Probability of Treatment by Distance from Peer Index Cut Point
Figure 1.3 – Test Scores by Distance from Peer Index Cut Point

A second condition for the validity of an RD design is that observations (in this case, schools) do not have precise control over the assignment variable (Lee and Lemieux, 2010). One test is to compare pre-treatment covariates to the forcing variable. This is shown for one important covariate – the average 2007 test score – in Figure 1.3. As in Figure 1.2, the x-axis shows the forcing variable – the peer index – grouped into bins of size 0.1. The left panel of Figure 1.3 restricts eligible schools to be those in the treatment group while the right panel restricts eligible schools to those in the control group. A line based on a local linear regression is
fit separately on either side of the cutoff and is superimposed on the scatterplot. For both groups and subjects, we can see a strong negative relationship between peer index and average math test scores prior to the introduction of the bonus program. The relationship appears to be fairly linear and, as we would expect, has no notable discontinuity at the cut point for bonus program eligibility.

1.4 Results

1.4.1. Student Test Scores

To test for results consistent with a loss of intrinsic motivation among teachers, I first look at student test scores. Table 1.4 shows the results of 30 separate regressions, each fit using Equation 1.1. In each regression, the dependent variable is a school’s average test score in math or ELA, measured in units of standard deviation. The sample of schools is restricted to be those that were eligible to for randomization into the bonus program. In the rows labeled “No Controls,” the regression is run on a constant and a single indicator variable that is equal to one if a school was randomly selected to be part of the treatment group and zero otherwise. In the rows labeled “With Controls,” additional independent variables are added to the regression to control for pre-treatment test scores and demographic variables. In all cases, the coefficients should be viewed as Intention to Treat (ITT) estimates as most (87%), but not all, invited schools elected to participate in the program (Marsh et al., 2011).

In Table 1.4, as expected, we see no impact of the bonus program on student test scores prior to the implementation of the program. In both 2006 and 2007, all coefficient estimates are close to zero, with three positive and three negative. These results serve as a placebo test and, in combination with Table 1.2, provide evidence that the randomization procedure was effective.

13 Results are weighted by the number of test takers in each school. Unweighted results are similar.
While adding controls to the regression in 2007 reduces the size of the standard error, it does not markedly change the size of the coefficients.\textsuperscript{14}

Table 1.4 – Impact of the Bonus Program on Average Student Test Scores

<table>
<thead>
<tr>
<th></th>
<th>Pre-Program</th>
<th>Program</th>
<th>Post-Program</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>$D_i$ Coefficient</td>
<td>2006</td>
<td>2007</td>
<td>2008</td>
</tr>
<tr>
<td>Math, No Controls</td>
<td>-0.03</td>
<td>0.00</td>
<td>-0.04</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.06)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>Math, With Controls</td>
<td>0.02</td>
<td>-0.04</td>
<td>-0.09**</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>ELA, No Controls</td>
<td>0.01</td>
<td>-0.02</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.06)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>ELA, With Controls</td>
<td>-0.02</td>
<td>-0.01</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>R-Squared</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math, No Controls</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Math, With Controls</td>
<td>0.77</td>
<td>0.78</td>
<td>0.63</td>
</tr>
<tr>
<td>ELA, No Controls</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ELA, With Controls</td>
<td>0.75</td>
<td>0.73</td>
<td>0.59</td>
</tr>
<tr>
<td>N</td>
<td>302</td>
<td>302</td>
<td>302</td>
</tr>
</tbody>
</table>

Note: each coefficient is the result of a separate regression with school-level mean test score as the dependent variable and treatment ($D_i$) as the key independent variable. Regressions with controls adjust for the average test score in the same subject in 2007 (or 2006 in the case when the 2007 test score is the dependent variable), the Progress Report Overall Score in 2007 (excluded in the case when the 2007 test score is the dependent variable), the Progress Report Peer Index in 2007, indicators for the school level (K8, ES, MS) in 2008, and the percentage of students in 2008 who were: English Language Learners, Special Education, eligible for free or reduced price lunches, and black or Hispanic. Regressions are weighted by the number of test takers in each school. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

During the treatment period – corresponding to the years 2008 through 2010 – we see a small but growing negative estimate of the teacher bonus program effect (often not statistically distinguishable from zero). In math, by 2010, treatment schools have test scores that are 0.12 standard deviations lower than control schools in the baseline specification and 0.13 standard deviations lower than control schools in the weighted specification.\textsuperscript{14}

\textsuperscript{14} It is not possible control for pre-treatment test scores in the 2006 regression because 2006 is the first year for which the NYCDOE provides test score data on its web site.
deviations lower in the specification with controls. The former estimate is significant at the 0.10 level and the latter at the 0.01 level. The increased statistical significance in the regression with controls is largely caused by a reduction in the standard error, as the coefficient changes little. In ELA, we see negative coefficients throughout the 2008 to 2010 time period that are not statistically distinguishable from zero. These results are broadly consistent with those of other researchers who have studied the effect of the NYC teacher bonus program on student outcomes, despite the fact that two studies (Marsh et. al., 2011; Fryer, 2013) use individual student data and one (Goodman and Turner, 2013) explores only the first two years of the program.

With respect to student test scores, one contribution of this paper is to extend the analysis beyond the conclusion of the teacher bonus program and look for results consistent with a loss of intrinsic motivation. In the last three columns of Table 1.4 – corresponding to the years 2011 to 2013 – we see such results. In math, the negative point estimates become larger in magnitude and more strongly statistically significant. In the uncontrolled regression, the math point estimate reaches a low point at -0.16 standard deviations in 2012. In ELA, the negative point estimates roughly double in magnitude between 2010 and 2011 and become statistically significant at the 0.05 level, whether or not one includes controls to improve precision. In the uncontrolled regression, the ELA point estimate reaches a low of -0.13 standard deviations in 2011, the year immediately following the conclusion of the bonus program.

These results are consistent with a story in which the introduction of the teacher bonus program reduced intrinsic motivation among teachers at the treatment schools. However, these results could be consistent with other potential explanations. Perhaps the bonus program caused teachers at treatment schools to change their practice in a way that was detrimental to student learning. Or, perhaps teachers at control schools – who may have been aware of the bonus
program through news reports – increased their productivity in an effort to compete with treatment schools and prove that merit pay doesn’t improve performance in New York City.

1.4.2 Teacher Survey Responses

To further explore whether the observed test score results are caused by a decline in teachers’ intrinsic motivation, I analyze teachers’ responses to the NYC School Survey. As discussed earlier, teachers’ responses were converted to Likert-like scale with 0 corresponding to the most negative response and 10 to the most positive. Each school received a score for each question in each year when that question was asked. Table 1.5 shows the impact of the bonus program on teacher responses to one sample question, where teachers were asked for their agreement with the following statement: “I usually look forward to each working day at my school.” It is fit using a simplified version of Equation 1.3, where \( z_{st} \) corresponds to school \( s \)’s score on the question in year \( t \). While this question was only asked in 2012 and 2013, two and three years after the completion of the bonus program, respectively, it is arguably the most direct measure of intrinsic motivation on the survey.

Table 1.5 – Impact of the Bonus Program on Teacher Agreement with the Statement: “I usually look forward to each working day at my school”

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2012</td>
<td>2013</td>
</tr>
<tr>
<td>( D_i ) Coefficient</td>
<td>-0.22</td>
<td>-0.04</td>
</tr>
<tr>
<td>(0.15)</td>
<td>(0.15)</td>
<td></td>
</tr>
<tr>
<td>( N )</td>
<td>297</td>
<td>301</td>
</tr>
</tbody>
</table>

Note: each coefficient is the result of a separate regression with teacher scores as the dependent variable and treatment \( (D_i) \) as the independent variable. Regressions are weighted by the number of responders at each school. Robust standard errors in parentheses. *** \( p<0.01 \), ** \( p<0.05 \), * \( p<0.1 \)
The coefficients are both negative – indicating that teachers at treatment schools were less likely to agree with this statement – but they are statistically indistinguishable from zero. For context, in 2012 this question had a mean score of 7.01 and a standard deviation of 1.22 in the control group. To better understand the trajectory of survey responses before, during, and after the bonus program, I focus on the 45 questions that were consistently asked on the teacher survey between 2007 and 2012. I divide these questions into those related and unrelated to intrinsic motivation in two ways. First, I subjectively divide them based on whether or not I consider the question to be related to intrinsic motivation (see Table A.1.1 in the appendix for a full list). Second, I conduct an exploratory factor analysis using data from the 2012 survey and identify those questions that measure the same underlying construct as the “look forward” question shown in Table 1.5.

Table 1.6 – Intrinsic Motivation Questions vs. Other Questions (Subjective Division)

<table>
<thead>
<tr>
<th></th>
<th>Pre-Pgm.</th>
<th>Program</th>
<th>Post-Program</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>$D_i$ Coefficient</td>
<td>2007</td>
<td>2008</td>
<td>2009</td>
</tr>
<tr>
<td>Intrinsic Motivation Questions</td>
<td>-0.16 (0.12)</td>
<td>-0.08 (0.11)</td>
<td>0.05 (0.12)</td>
</tr>
<tr>
<td>Other Questions</td>
<td>-0.15 (0.13)</td>
<td>-0.06 (0.11)</td>
<td>0.03 (0.12)</td>
</tr>
<tr>
<td>Difference (Intrinsic - Other)</td>
<td>-0.01 (0.06)</td>
<td>-0.02 (0.05)</td>
<td>0.02 (0.04)</td>
</tr>
<tr>
<td>N</td>
<td>299</td>
<td>302</td>
<td>302</td>
</tr>
</tbody>
</table>

Note: each coefficient is the result of a separate regression of the mean school-level teacher survey scores as the dependent variable and treatment ($D_i$) as the independent variable. Regressions are weighted by the number of responders at each school. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1
Tables 1.6 shows the results when questions are subjectively divided into those that are, and are not, related to intrinsic motivation. In the top row, we see that the treatment group had intrinsic motivation survey scores similar to the control group both before and during the bonus program. In 2011 and 2012, after the bonus program concluded, the intrinsic motivation survey scores for the treatment group were between 0.23 and 0.24 points lower than for the control group, a result that is statistically significant at the 0.05 level. This fact, taken alone, is consistent with a reduction in teachers’ intrinsic motivation at schools invited to participate in the bonus program. However, in the second row of the table, we see a very similar pattern emerge among questions that are unrelated to intrinsic motivation. In the third row, we see results from a difference-in-differences regression where the first difference is between the treatment and control group and the second difference is between those questions related to intrinsic motivation and those that are unrelated.\textsuperscript{15} In this row, we see precisely estimated zeros, indicating that it is unlikely that the bonus program had a greater effect on teacher responses to intrinsic motivation questions than on teacher responses to other questions.

In Table 1.7, I show a similar set of regressions coefficients where the survey questions have been divided based on the results of an exploratory factor analysis. Those labeled “Intrinsic Motivation Questions” in Table 1.5 are a set of 18 questions that loaded onto the same factor as the “look forward” question discussed above. As a comparison group, in row 2, I select a group of six questions that load onto a factor that appears to be related to school safety.\textsuperscript{16} The results in Table 1.7 are very similar to those in Table 1.6, though the coefficients are less precisely

\textsuperscript{15} This specification is not a standard difference-in-differences regression, since neither difference is related to time. In the appendix, however, I take time into account in two ways. First, I estimate a model that controls for pre-program scores in 2007. Second, I use a triple-difference model where the first difference is between treatment and control, the second difference is between intrinsic motivation questions and non-intrinsic motivation questions, and the third difference is between each year and 2007. Once 2007 is controlled for, the results in rows 1 and 2 are similar in direction but no longer statistically significant at the 0.05 level. The results in row 3 continue to be precisely-estimated zeros.

\textsuperscript{16} For example, one survey item in this group is: “Gang activity is a problem in my school.”
estimated. In both tables, it appears that the participating in the bonus program may have had a negative effect on teacher survey scores in 2011 and 2012. However there is no indication that intrinsic motivation was more affected than other attitudes measured on the survey.

Table 1.7 – Intrinsic Motivation Questions vs. Other Questions (Factor Analysis Division)

<table>
<thead>
<tr>
<th></th>
<th>Pre-Pgm.</th>
<th>Program</th>
<th>Post-Program</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td></td>
<td>2007</td>
<td>2008</td>
<td>2009</td>
</tr>
<tr>
<td>Intrinsic Motivation Questions</td>
<td>-0.20</td>
<td>-0.05</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>(0.16)</td>
<td>(0.13)</td>
<td>(0.14)</td>
</tr>
<tr>
<td>School Safety Questions</td>
<td>-0.19</td>
<td>-0.16</td>
<td>-0.08</td>
</tr>
<tr>
<td></td>
<td>(0.15)</td>
<td>(0.13)</td>
<td>(0.13)</td>
</tr>
<tr>
<td>Difference (Intrinsic - Safety)</td>
<td>-0.01</td>
<td>0.10</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>(0.16)</td>
<td>(0.13)</td>
<td>(0.11)</td>
</tr>
<tr>
<td>N</td>
<td>299</td>
<td>302</td>
<td>302</td>
</tr>
</tbody>
</table>

Note: each coefficient is the result of a separate regression of the mean school-level teacher survey scores as the dependent variable and treatment \((D_i)\) as the independent variable. Regressions are weighted by the number of responders at each school. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

1.4.3 Regression Discontinuity Evidence

In the test score and survey results shown above, I present the estimated coefficients on a \textit{treatment} indicator – equal to 1 for treatment schools and 0 for control schools – rather than on a \textit{control} indicator. Doing so implicitly suggests that results at the treatment schools declined relative to their counterfactual. However, it is also plausible that the treatment schools held steady – on the path they would have had the bonus program not been implemented – and the control schools improved. To test for such a phenomenon, which would be inconsistent with a decline in intrinsic motivation among treatment teachers, I use a Regression Discontinuity (RD)
To implement the RD study, I take advantage of the rigid discontinuity in school eligibility for the bonus program. Prior to randomization, 430 high-need schools were selected to be eligible for the bonus based on their peer index on the NYC Progress Report. 17 These *bonus-eligible* schools – a subset of all 1,217 NYC schools that received a Progress Report in 2007 – were selected entirely based on their Progress Report peer index. Within each of five school types – elementary, K-8, middle, high, and transfer – schools above a particular peer index score were chosen to be eligible for the teacher bonus program and schools below that score were ineligible. 18 Conceptually, for an RD study, I want to compare bonus-eligible schools just above the peer index cut point to bonus-ineligible schools just below the cut point. Assuming that bonus-ineligible schools just below the cut point present a valid counterfactual for bonus-eligible schools just above the cut point, once the peer index (i.e. forcing variable) is controlled for, any difference in outcomes reflects the impact of being eligible for the bonus program.

While the 302 elementary, middle, and K-8 schools I use in my main analysis are valid for an RCT, there are two adjustments I need to make to ensure the validity of the RD. First, as documented in Fryer (2013), prior to randomization, 34 bonus-eligible schools were barred by the teachers’ union (UFT) for “unknown reasons.” I add back the 15 elementary, middle, and K-8 schools among these 34 to ensure that any effect found by the RD is not caused by the UFT excluding these schools from the eligible group (since they made no similar exclusion among the

---

17 Note that my study focuses on the 302 elementary, middle, and K-8 schools that had consistent testing data from 2007 to 2013 and were not barred from participation by the UFT.
18 For middle schools, high schools, and transfer schools, a lower peer index meant a higher-need school. Technically, therefore, schools *below* a certain peer index threshold were eligible for the bonus program. In this analysis, I normalize all peer indices by calculating z-scores within school type and then defining a higher z-score to be a higher-need schools (multiplying by -1 where needed).
ineligible group). Second, schools that spanned the middle and high school grades had two chances to become bonus-eligible, since they received a separate peer index for their middle and high school portions, and qualified if either peer index was above the cutoff. Prior to the RD, I remove all schools that received both middle and high school peer indices from the analysis, whether or not they appeared in the bonus-eligible group. This reduced the bonus-eligible group by six schools and the non-eligible group by 24 schools. In total, this leaves 311 bonus-eligible elementary, middle, and K-8 schools which I compare to 607 non-eligible schools.

Figure 1.4 – Test Scores by Distance from Peer Index Cut Point (2011)
Figure 1.4 presents the results of the RD graphically for 2011, the first year after the bonus program was concluded. The graph is organized in the same way as Figure 1.3, with math test scores on the top and ELA test scores on the bottom. In all four panels, to the left of the dotted line we see average test scores for schools that were not eligible for randomization based on their peer index. These schools serve as the counterfactual for bonus-eligible schools. In the left two panels, which compare non-eligible schools to treatment schools, we do not see a notable discontinuity in 2011 test scores between the two groups. To the extent there is any discontinuity, it appears that treatment schools may have outperformed non-eligible schools. In the right two panels, which compare non-eligible schools to control schools, we see a more noticeable discontinuity. In particular, control schools near the cut point – just to the right of the dotted line – appear to have higher 2011 average test scores than would be predicted based on their peer index.

Table 1.8 provides regression results to support the visual evidence in Figure 1.4. Each entry in Table 1.8 shows the coefficient and standard error for $E_s$ – a dummy variable indicating eligibility to be randomized for the teacher bonus program – from a linear regression to fit Equation 1.4. As explained above, I fit a linear regression model because the data – even in the pre-program period – have a strong linear relationship and higher-order polynomials appear to overfit the data, failing the placebo tests. Consistent with the recommendation of Gelman and Imbens (2014), I also present RD results using a local linear regression in an appendix, and the results are very similar to those presented in Table 1.8.

In Table 1.8, we see that eligibility for the bonus program had no impact on test scores in either the treatment or control group in the two years prior to the program’s announcement (2006
This result, consistent with the visual evidence in Figure 1.3, serves as a falsification test. In the top two rows of Table 1.8, we see little evidence of an effect of the bonus program on treatment school test scores after the program began in 2008. The treatment-school estimates are generally fairly small and not significantly different from zero. If anything, the coefficients are somewhat more likely to be positive than negative, a finding that is inconsistent with what one would expect if teachers’ intrinsic motivation had declined.

Table 1.8 – Impact of Bonus Eligibility on Student Test Scores (Regression Discontinuity)

<table>
<thead>
<tr>
<th>$E_i$</th>
<th>Pre-Program</th>
<th>Program</th>
<th>Post-Program</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) (2) (3) (4) (5) (6) (7) (8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2006 2007 2008 2009 2010 2011 2012 2013</td>
<td>N = 793 793 745 745</td>
<td></td>
</tr>
<tr>
<td>Math, Treatment</td>
<td>-0.01 0.00 -0.01 0.08 0.06 0.12 0.07 -0.05</td>
<td>(0.07) (0.07) (0.07) (0.09) (0.09) (0.09) (0.09) (0.09)</td>
<td></td>
</tr>
<tr>
<td>ELA, Treatment</td>
<td>0.01 0.03 0.08 0.08 0.14* 0.11 0.02 -0.04</td>
<td>(0.07) (0.06) (0.07) (0.07) (0.07) (0.08) (0.08) (0.08)</td>
<td></td>
</tr>
<tr>
<td>Math, Control</td>
<td>0.05 0.05 0.06 0.16* 0.17* 0.28*** 0.24** 0.09</td>
<td>(0.07) (0.08) (0.08) (0.09) (0.09) (0.10) (0.09) (0.09)</td>
<td></td>
</tr>
<tr>
<td>ELA, Control</td>
<td>0.08 0.07 0.15** 0.16** 0.20*** 0.21*** 0.15** 0.08</td>
<td>(0.07) (0.07) (0.07) (0.07) (0.07) (0.08) (0.07) (0.08)</td>
<td></td>
</tr>
</tbody>
</table>

Note: Each number is the coefficient on $E_i$ in a separate regression of school-level mean test score on distance from the eligibility cut point, an indicator for bonus eligibility ($E_i$), and an interaction of the two (allowing the slopes to vary on either side of the cut point). Regressions are weighted by the number of test takers in each school. Robust standard errors are in parentheses. The sample size is the same in each year because only schools with testing data in each year were included. *** p<0.01, ** p<0.05, * p<0.1

The bottom two rows of Table 1.8 restrict the eligible sample to control schools only. When running the RD, I am then comparing ineligible schools – which never had the opportunity to be randomized into the bonus program – with eligible schools that were randomly selected to be control schools. In Table 1.8, we see that control schools right around the eligibility cut point had significantly higher test scores than ineligible schools with similar peer indices. Between
2009 and 2012, the estimated coefficients range from 0.15 to 0.28 school-level standard deviations, and all results are significant at the 0.10 level with most significant at the 0.05 level. These results – when combined with the null effects observed in treatment schools – provide more nuance to the negative RCT results highlighted earlier in this paper. It appears that the negative impact of the bonus program was not so much the result of a decline in performance among treatment schools, but rather of an improvement in performance among control schools.

Table 1.9 – Comparison of RD and RCT Results

<table>
<thead>
<tr>
<th></th>
<th>Pre-Program</th>
<th>Program</th>
<th>Post-Program</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. RD (Treatment Schools)</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>Math (All Schools)</td>
<td>-0.01</td>
<td>0.00</td>
<td>-0.01</td>
</tr>
<tr>
<td>ELA (All Schools)</td>
<td>0.01</td>
<td>0.03</td>
<td>0.08</td>
</tr>
<tr>
<td>B. RD (Control Schools)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math (All Schools)</td>
<td>0.05</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td>ELA (All Schools)</td>
<td>0.08</td>
<td>0.07</td>
<td>0.15**</td>
</tr>
<tr>
<td>C. Treatment - Control</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math (All Schools)</td>
<td>-0.06</td>
<td>-0.05</td>
<td>-0.07</td>
</tr>
<tr>
<td>ELA (All Schools)</td>
<td>-0.07</td>
<td>-0.04</td>
<td>-0.07</td>
</tr>
<tr>
<td>D. RCT Results (Treatment Effect)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math, No Controls</td>
<td>-0.03</td>
<td>0.00</td>
<td>-0.04</td>
</tr>
<tr>
<td>ELA, No Controls</td>
<td>0.01</td>
<td>-0.02</td>
<td>-0.02</td>
</tr>
</tbody>
</table>

Sample Size

<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A (Treatment)</td>
</tr>
<tr>
<td>Panel B (Control)</td>
</tr>
</tbody>
</table>

Note: Each number in Panel A and B is the coefficient on a separate regression of school-level mean test score on distance from the eligibility cut point, an indicator for bonus eligibility (Ei), and an interaction of the two (allowing the slopes to vary on either side of the cut point). Panel A restricts eligible schools to include treatment schools only. Panel B restricts eligible schools to include control schools only. Panel C shows the difference between the estimates in Panel A and Panel B. Panel D reproduces the treatment coefficient from the "no control" regressions in Table 2. Regressions are weighted by the number of test takers in each school. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1
Table 1.9 shows the degree to which the RD and RCT results align with one another. In Panel A and B, I replicate the coefficients from Table 1.8, omitting standard errors for clarity. In Panel C, I subtract the RD coefficients for control schools from the RD coefficients for treatment schools, obtaining one estimate of the difference between treatment and control schools. These estimates are very close to the RCT results from Table 1.4, which I have replicated (for the regression with no controls) in Panel D of Table 1.9. While the RCT and the RD show largely the same estimated difference between the treatment and control group, the RD gives us more insight into the cause of this difference. Rather than a decline in performance among treated schools – as would be consistent with a decline in teachers’ intrinsic motivation – it appears likely that control schools improved their performance over what would have happened in the absence of the bonus program.

1.5 Conclusion

In this paper, I test whether the implementation of a large teacher bonus program in New York City had a detrimental effect on teachers’ intrinsic motivation. Initially reviewing data from the Randomized Controlled Trial, I find some evidence consistent with this hypothesis. Teachers at treatment schools responded more negatively to questions related to intrinsic motivation and students at treatment schools scored lower on standardized tests than their counterparts at randomly assigned control schools. However, exploring further, it appears unlikely that a decline in teachers’ intrinsic motivation drove these results. When looking at teacher survey scores, we see that scores at treatment schools dropped by a similar amount for questions unrelated to intrinsic motivation. And when looking at student test scores, a Regression Discontinuity analysis shows that students at treatment schools did not, in fact,
decline relative to their (unobserved) counterfactual, as would be predicted if the bonus program lowered teachers’ intrinsic motivation. Rather, it appears that students at control schools scored higher than their (also unobserved) counterfactual.

While this finding is unexpected, it is not infeasible. Given the public nature of the bonus program, it is possible that teachers at control schools knew that they were working in control schools and made an effort to improve their performance. The bonus program was widely publicized when it was implemented, and the NYCDOE published a full list of invited treatment schools that, if combined with information on the peer index, would have enabled schools to know if they were in the control group with reasonable certainty.\textsuperscript{19} \textsuperscript{20} Moreover, it has been widely observed that the behavior of experimental units may change once they are aware they are participating in a study, a phenomenon broadly referred to as a Hawthorne Effect. Once specific form is known as the “John Henry Effect,” in which experimental participants are aware that they are in the control group and seek to work harder as a result. Such an effect, if it took place during the NYC teacher bonus program, might explain the results observed in this paper.

More broadly, the RD results serve as an important reminder when interpreting RCT results. An RCT is generally considered the “gold standard” for causal evidence in the social sciences because the control group is assumed to be a valid counterfactual for the treatment group. The control group answers the question: what would have happened to the treatment group had the experiment not be implemented? However, the results in this paper remind us that, even in an RCT, the treatment group counterfactual is never actually observed. Even with proper randomization, as by all accounts occurred in the NYC teacher bonus program, the

\textsuperscript{19} See, for example, a New York Times article in 2008 announcing the distribution of $14.2 million dollars to teachers after the first year of the program (http://www.nytimes.com/2008/09/19/education/19bonus.html)

\textsuperscript{20} The announcement with the list of invited treatment schools was available at the following link until recently: http://schools.nyc.gov/Offices/mediarelations/NewsandSpeeches/2007-2008/20071218_performance_pay.htm. I have an archived copy and can provide it upon request.
control group is still an imperfect estimate of what would have happened in the treatment group had the experiment not occurred.
2. Moral Hazard and Less Invasive Medical Treatment for Coronary Artery Disease: The Case of Cigarette Smoking

Jesse Margolis, Jason Hockenberry, Michael Grossman, and Shin-Yi Chou
2.1 Introduction

Coronary Artery Disease (CAD) is a common and deadly disease. In 2010, over 350,000 people died of CAD in the United States, making the disease responsible for roughly one in seven deaths (Murphy, Xu, and Kochanek, 2012). CAD is caused by a buildup of plaque on the arterial walls leading to the heart, resulting in reduced blood flow. If the buildup is not checked, CAD can result in an acute myocardial infarction (AMI, a.k.a. “heart attack”) due to insufficient oxygen reaching the heart.

A number of medical treatments are available to patients with CAD. First, and least invasive, is “medical management.” Medical management involves non-surgical treatment including prescription medication, lifestyle modification, and frequent monitoring. The second treatment is a revascularization procedure known as Percutaneous Coronary Intervention (PCI, commonly referred to as angioplasty). A doctor (usually an interventional cardiologist) performing PCI makes a small incision and arthroscopically inserts and inflates a balloon at the site of the lesion to expand the vessel. PCI in the modern era usually involves the placement of a wire mesh stent at the blockage site, which assists in keeping the arterial walls expanded to maintain blood flow. The PCI procedure takes approximately 60 minutes and the patient usually

---

21 Research for this paper was supported by grant number 5R21 AG033876 from the National Institute on Aging and the Office of Behavioral and Social Sciences Research to the National Bureau of Economic Research. This paper was presented at seminars at the University of Chicago, the University of Illinois at Chicago, Johns Hopkins University, the University of Melbourne, the Ohio State University, the University of Connecticut, the University of Delaware, the Weill Cornell Medical School, Columbia University, and Washington State University. It also was presented at the 2013 Bureau of the Census Research Data Centers Conference at the Federal Reserve Bank of Atlanta, at the 2014 spring meeting of the National Bureau of Economic Research Health Economics Program, and at the Fifth Biennial Conference of the American Society of Health Economists at the University of Southern California in June 2014. We would like to thank the participants in those forums for helpful comments and suggestions. We also would like to thank Peter Cram and John O’Grady for their medical insights and Sandra Decker for her advice on working with the linked NHIS/Medicare data. We are indebted to Jonathan Fisher, Research Data Center (RDC) Administrator at the New York Bureau of the Census RDC at Baruch College, and Frances McCarty, Senior Service Fellow at the National Center for Health Statistics RDC, for their assistance in helping us to gain access to and work with restricted files from the National Health Interview Survey.
spends one night in the hospital.\textsuperscript{22} The third and generally most invasive treatment is Coronary Artery Bypass Graft (CABG)\textsuperscript{23}, a major surgical procedure that involves harvesting a section of vessel from a different area of the body (either vessels in the groin or chest wall), opening the chest cavity via a sternectomy, and connecting one healthy part of the diseased artery to another, surgically bypassing the lesion. CABG surgery takes approximately four hours and patients generally spend at least a week recovering in the hospital.\textsuperscript{24}

Of the two procedures, PCI was developed more recently, having been initially used in the late 1970s, more than a decade after CABG was first performed. Its use expanded rapidly upon FDA approval of the coronary stent in 1994 (Cutler and Huckman, 2003). By 2010, the CDC estimated that there were 954,000 PCI procedures and 395,000 CABG procedures performed in the United States, with roughly half of them performed in patients over 65 (CDC, 2010).\textsuperscript{25}

Since the development of PCI, there have been numerous studies comparing the effectiveness of the two procedures in various populations (see Hlatky et al., 2009, Weintraub et al., 2012, and Mohr et al., 2013 for three recent studies that summarize prior research). While the results vary, the emerging consensus is that CABG patients have worse short-run outcomes than similar PCI patients – partly due to higher perioperative mortality – but better long-term outcomes. Results from a large observational study (Weintraub et al., 2012) are reproduced in Figure 2.1, showing that the survival curve for CABG patients is initially lower than for PCI patients, but that this trend is reversed a year or more after the procedure. A meta-analysis of ten

\textsuperscript{22} http://www.medicinenet.com/coronary_angioplasty/article.htm (accessed 3/12/15)
\textsuperscript{23} Less invasive CABG procedures have been in development and increasing use in recent years, though these were very infrequent during the period we examine.
\textsuperscript{24} http://www.medicinenet.com/coronary_artery_bypass_graft/article.htm (accessed 3/12/15)
\textsuperscript{25} These counts are based on ICD-9-CM procedure codes beginning with 361 for CABG and 0066, 3606, and 3607 for PCI. Note that the PCI codes used by the CDC to estimate the number of procedures are slightly different than those we use to identify PCI patients in our main analysis.
randomized controlled trials shows a similar pattern (Hlatky et al., 2009).

Figure 2.1 – Comparative Effectiveness: PCI vs. CABG (Weintraub et al., 2012)

While studies show that otherwise similar CABG patients have better long run outcomes than PCI patients, there is little consensus as to why.26 In this paper, we propose a novel hypothesis: patient offsetting behavior. Specifically, a patient who undergoes CABG rather than PCI is more likely to change her behavior in a way that promotes good health and a longer life: she is more likely to quit smoking, begin exercising, improve her diet, and avoid excessive alcohol intake. This might happen because the more invasive nature of CABG – a patient’s heart

26 Most authors either omit a discussion of mechanisms entirely (e.g. Weintraub et al., 2012) or imply that CABG may have mechanical benefits in the long run. For example, in commenting on the superiority of CABG in the recent SYNTAX randomized trial, Taggart (2013, p. 606) states that “… CABG and PCI achieve their benefits through quite different pathophysiological effects. Pathologically, most coronary artery disease is located in the proximal coronary arteries and bypass grafts to the mid-coronary vessels not only make the complexity of proximal disease irrelevant but also offer prophylaxis against the development of de-novo proximal disease. By contrast, although PCI can be highly effective in directly treating less complex proximal coronary artery disease, its benefits are mitigated by the development of new disease proximal to, within, or immediately distal to the stent…”

40
and lungs are bypassed during the surgery, she is in the hospital for a week, has a longer postoperative recovery period, and is left with a major scar and residual pain from the sternectomy – sends a stronger signal to the patient that she has a serious health problem.

This hypothesis is consistent with a prior economic research on moral hazard, showing that individuals change their behavior when their perceived risks change. Peltzman’s (1975) study of the effects of automobile safety regulation is a classic and seminal example. He develops a model in which the legal mandate to install various safety devices on automobiles lowers the price of fast and reckless driving because it lowers the probability that the driver will die in an accident. Hence, the demand for this activity rises. Empirically, he finds that the increase in this offsetting behavior (reckless driving) is so large that the regulations at issue had little impact on highway deaths and actually increased pedestrian deaths. More recently, Dave and Kaestner (2009) investigate the impact of health insurance access on health behaviors of the elderly, showing that access to Medicare at age 65 leads to a reduction in preventative behaviors and an increase in risky health behavior amongst the elderly. Peltzman (2011) demonstrates how medical technology breakthroughs can lead to offsetting behavior by showing that the age cohorts that benefited the most from the introduction of antibiotics experienced worse mortality rates from risky health behaviors. Kaestner, Darden, and Lakdawalla (2014) find that the use of statins leads to a small increase in body mass index and moderate (20–33 percent) increases in the probability of being obese, possibly because it changes the user’s perceived risk of consuming high calorie fattening foods.

In this study, we test one potential behavioral response to surgery – smoking – and see results consistent with patient offsetting behavior. Patients who undergo CABG – the more invasive procedure – are 12 percentage points more likely to quit smoking than PCI patients.
Our results are robust to a number of different specifications and, for reasons we discuss below, unlikely to be driven by selection bias.

2.2 Data

We use individual Medicare data merged with responses from the National Health Interview Survey (NHIS). The Medicare records identify those patients who have been diagnosed with CAD and show which of them have undergone PCI or CABG, along with the exact date of each diagnoses and procedure. The Medicare data also allow us to control for disease severity and other conditions that might be correlated with procedure type and induce quitting, such as a myocardial infarction (a.k.a. “heart attack”). The NHIS provides information on smoking and quitting behavior, as well as individual characteristics.

The Medicare data are provided by the Center for Medicare and Medicaid Services (CMS). To identify CAD patients and the type of treatment they underwent, we use the Medicare Standard Analytical Files, including the Inpatient, Outpatient, Skilled Nursing Facility, Carrier, Durable Medical Equipment, Home Health Agency, and Hospice claims files. These files contain one or more records for each individual. Each record contains the ICD-9-CM codes for all diagnoses made and procedures performed during that stay or claim. We identify CAD patients as those who have at least one diagnosis code beginning with 410, 411, 412, 413, or 414. We identify PCI patients as those CAD patients with at least one procedure code beginning with 0066, 3601, 3602, 3605, or 3606. We identify CABG patients as those CAD

---

27 A single record in the Inpatient file corresponds to a stay in a hospital. A single record in the Skilled Nursing Facility file corresponds to a stay in a Skilled Nursing Facility. A single record in the Outpatient file corresponds to a claim by an institutional outpatient provider (Hospital outpatient clinic, rural health clinics, etc.). A single record in the Carrier claim file corresponds to a claim by a non-institutional outpatient provider (physicians, physician assistants, etc.)
patients with procedure codes beginning with 361.\(^\text{28}\) Finally, we identify medically managed patients as those patients who have been diagnosed with CAD, but do not have a concurrent or subsequent PCI or CABG procedure.\(^\text{29}\)

The NHIS is an annual survey of approximately 85,000 individuals in over 30,000 U.S. households run by the National Center for Health Statistics (NCHS), part of the Centers for Disease Control and Prevention (CDC). All participants are asked questions about their general state of health and disability. Each year, a subset of approximately 30,000 individuals is asked about their smoking habits. These respondents are asked if they have ever smoked 100 cigarettes in their life. For those who say yes, they are asked if they currently smoke every day, some days, or not at all. If they do not currently smoke, they are asked when they quit, a question they can answer in days, weeks, months, or years. We use the responses to these questions to create a synthetic panel, identifying whether a person smoked on each date prior to their NHIS interview. Each person is categorized as either an always smoker, a never smoker, or a quitter who smoked up to the day she reports quitting.\(^\text{30}\)

The individual NHIS responses have been linked to Medicare data by the CDC and CMS and made available as a restricted-use dataset to researchers. The linkage is based on social security number, date of birth, and gender. To be linked, the data must match on all three fields. To date, the CDC and CMS have linked the 1994-1998 NHIS surveys to Medicare data from 1991-2007 and the 1999-2005 NHIS surveys to Medicare data from 1999-2007. The linkage is described further in the appendix.

---

\(^{28}\) For both PCI and CABG, we exclude the small number of patients who do not have a concurrent or prior CAD diagnosis.

\(^{29}\) A patient who is diagnosed with CAD before her NHIS interview date and has PCI or CABG after her NHIS interview date is counted as medically managed at the time of the NHIS interview.

\(^{30}\) This categorization vastly over-simplifies the complexity of smoking and quitting behavior, but still allows us to investigate our key question: what is the difference in quitting behavior between CAD patients undergoing medical management, PCI, and CABG.
2.3 Initial Analysis

In total, 12,265 NHIS respondents were linked to Medicare data and diagnosed with CAD.\(^{31}\) Of these individuals, between the date of their diagnosis and the date of their NHIS interview, 10,713 patients were treated only with medical management, 771 patients underwent PCI but not CABG surgery, and 781 patients underwent CABG surgery.\(^{32}\) Though our focus is on the two procedures – PCI and CABG – we include medically managed patients in all analyses for two reasons. First, due to the substantially greater number of medically managed patients, including them improves the precision of our covariate estimates (e.g. determining the impact of having a heart attack on smoking). Second, since medical management is the least invasive treatment for CAD, we might expect medically managed patients to quit at the lowest rate. This result, which we find in the data, supports our theory that changes in smoking behavior are related to treatment invasiveness. Our main finding, however – that CABG patients quit smoking at a higher rate than PCI patients – is robust to excluding medically managed patients from the analysis entirely.

Basic characteristics of the 12,265 CAD patients are shown in Table 2.1.\(^{33}\) Overall, when compared to patients undergoing medical management, patients who undergo a procedure (PCI or CABG) are more likely to be younger, male, and white. PCI and CABG patients appear to have largely similar demographic characteristics, though CABG patients are somewhat more likely to be male. When comparing medical conditions, both PCI and CABG patients are substantially more likely than medically managed patients to have had their first Acute

---

\(^{31}\) To be included, patients had to be diagnosed with CAD after the start of our Medicare data, but before the date of their NHIS interview (so that we have information on their smoking behavior both before and after their treatment).\(^{32}\) These counts are weighted by the NHIS probability weights. The unweighted totals are 10,772 medically managed patients, 723 PCI patients, and 770 CABG patients. Unweighted, 99 patients underwent both PCI and CABG surgery. These patients are included in the CABG category, because that is the more invasive treatment. Our results are robust to including them in the PCI category or excluding them altogether.\(^{33}\) Results presented in this paper include all Medicare participants, regardless of age. Results excluding those under 65, available upon request, are similar.
Myocardial Infarction (AMI, a.k.a. “heart attack”) within six months of initiating treatment. A number of other comorbidities – including congestive heart failure and valvular disease – show up most frequently in CABG patients, followed by PCI patients. In some of our regression specifications, we control for the covariates shown in Table 2.1.

### Table 2.1 – Patient Characteristics by Treatment

<table>
<thead>
<tr>
<th>Demographic Characteristics</th>
<th>MM</th>
<th>PCI</th>
<th>CABG</th>
<th>Medical Conditions</th>
<th>MM</th>
<th>PCI</th>
<th>CABG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td><strong>First AMI Within 6 Months of Treatment</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 55</td>
<td>4%</td>
<td>4%</td>
<td>2%</td>
<td>Yes</td>
<td>8%</td>
<td>40%</td>
<td>37%</td>
</tr>
<tr>
<td>55-64</td>
<td>7%</td>
<td>8%</td>
<td>7%</td>
<td>No</td>
<td>92%</td>
<td>60%</td>
<td>63%</td>
</tr>
<tr>
<td>65-69</td>
<td>24%</td>
<td>25%</td>
<td>26%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>70-74</td>
<td>22%</td>
<td>25%</td>
<td>26%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75-79</td>
<td>21%</td>
<td>20%</td>
<td>23%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80-84</td>
<td>13%</td>
<td>13%</td>
<td>13%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>85+</td>
<td>8%</td>
<td>5%</td>
<td>2%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>% With Comorbidity Within 6 Months of Treatment</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>44%</td>
<td>53%</td>
<td>59%</td>
<td>Congestive heart failure</td>
<td>14%</td>
<td>21%</td>
<td>34%</td>
</tr>
<tr>
<td>Female</td>
<td>56%</td>
<td>47%</td>
<td>41%</td>
<td>Pulmonary disease</td>
<td>11%</td>
<td>19%</td>
<td>27%</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td>Other neurological</td>
<td>3%</td>
<td>3%</td>
<td>3%</td>
</tr>
<tr>
<td>Asian</td>
<td>1%</td>
<td>1%</td>
<td>1%</td>
<td>Chronic pulmonary disease</td>
<td>15%</td>
<td>12%</td>
<td>22%</td>
</tr>
<tr>
<td>Black</td>
<td>9%</td>
<td>6%</td>
<td>4%</td>
<td>Diabetes w/o chronic comp.</td>
<td>14%</td>
<td>16%</td>
<td>17%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>5%</td>
<td>3%</td>
<td>4%</td>
<td>Diabetes w/ chronic comp.</td>
<td>4%</td>
<td>7%</td>
<td>10%</td>
</tr>
<tr>
<td>White</td>
<td>85%</td>
<td>89%</td>
<td>90%</td>
<td>Hypothyroidism</td>
<td>7%</td>
<td>6%</td>
<td>5%</td>
</tr>
<tr>
<td>Mult./Oth/Unknown</td>
<td>1%</td>
<td>1%</td>
<td>1%</td>
<td>Renal failure</td>
<td>2%</td>
<td>4%</td>
<td>4%</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
<td></td>
<td>Liver disease</td>
<td>1%</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td>Elem (K-8)</td>
<td>21%</td>
<td>17%</td>
<td>20%</td>
<td>Chronic Peptic ulcer disease</td>
<td>0%</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td>HS (non-grad); GED</td>
<td>19%</td>
<td>23%</td>
<td>18%</td>
<td>HIV and AIDS</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>HS grad</td>
<td>29%</td>
<td>30%</td>
<td>29%</td>
<td>Lymphoma</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Some col; AA deg.</td>
<td>18%</td>
<td>20%</td>
<td>18%</td>
<td>Metastatic cancer</td>
<td>1%</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>BA degree</td>
<td>7%</td>
<td>7%</td>
<td>8%</td>
<td>Solid tumor without metastasis</td>
<td>5%</td>
<td>5%</td>
<td>3%</td>
</tr>
<tr>
<td>Grad. Degree</td>
<td>5%</td>
<td>5%</td>
<td>6%</td>
<td>Rheumatoid arthritis</td>
<td>3%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>Unknown</td>
<td>1%</td>
<td>0%</td>
<td>1%</td>
<td>Coagulation deficiency</td>
<td>3%</td>
<td>4%</td>
<td>9%</td>
</tr>
<tr>
<td><strong>Family Income</strong></td>
<td></td>
<td></td>
<td></td>
<td>Obesity</td>
<td>3%</td>
<td>8%</td>
<td>9%</td>
</tr>
<tr>
<td>$0 to $9,999</td>
<td>19%</td>
<td>15%</td>
<td>14%</td>
<td>Weight loss</td>
<td>1%</td>
<td>1%</td>
<td>2%</td>
</tr>
<tr>
<td>$10,000 to $19,999</td>
<td>25%</td>
<td>22%</td>
<td>24%</td>
<td>Fluid and electrolyte disorders</td>
<td>10%</td>
<td>13%</td>
<td>24%</td>
</tr>
<tr>
<td>$20,000 to $35,000</td>
<td>20%</td>
<td>24%</td>
<td>24%</td>
<td>Blood loss anemia</td>
<td>1%</td>
<td>3%</td>
<td>3%</td>
</tr>
<tr>
<td>$35,000 or over</td>
<td>18%</td>
<td>22%</td>
<td>21%</td>
<td>Deficiency anemias</td>
<td>11%</td>
<td>14%</td>
<td>23%</td>
</tr>
<tr>
<td>Unknown</td>
<td>18%</td>
<td>17%</td>
<td>17%</td>
<td>Alcohol abuse</td>
<td>1%</td>
<td>1%</td>
<td>1%</td>
</tr>
</tbody>
</table>
| **Note:** All data are weighted by the NHIS probability weights. Age is as of diagnosis (CAD) or procedure (PCI / CABG). Comorbidities based on Elixhauser et al. (1998). * Within six months on either side of treatment (i.e. a one-year window).
Table 2.2 shows the smoking status of each group of respondents – medical management, PCI, and CABG – as of the date of the NHIS interview. Two items merit notice. First, CABG and PCI patients are more likely to have ever smoked than medically managed patients (i.e. the percentage of respondents who never smoked is lower for CABG and PCI patients). Second, most people who have ever smoked have quit smoking by the time of the NHIS interview, a trend that is most pronounced for CABG patients. While 61.0 percent of CABG patients in our study smoked at some point in their life, only 9.1 percent smoke as of their NHIS interview. PCI patients have a lower proportion of quitters, followed by medically managed patients.

Table 2.2 – Smoking Status as of NHIS Interview Date

<table>
<thead>
<tr>
<th>Smoking Status as of Survey</th>
<th>Treatment</th>
<th>MM</th>
<th>PCI</th>
<th>CABG</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current</td>
<td></td>
<td>12.2%</td>
<td>11.3%</td>
<td>9.1%</td>
<td>11.9%</td>
</tr>
<tr>
<td>Quit</td>
<td></td>
<td>42.6%</td>
<td>49.3%</td>
<td>52.0%</td>
<td>43.6%</td>
</tr>
<tr>
<td>Never Smoked</td>
<td></td>
<td>45.3%</td>
<td>39.5%</td>
<td>39.0%</td>
<td>44.5%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Count</td>
<td></td>
<td>10,713</td>
<td>771</td>
<td>781</td>
<td>12,265</td>
</tr>
</tbody>
</table>

Note: This table shows the smoking status of every NHIS respondent who was diagnosed with CAD prior to their interview date. Data are weighted by the NHIS probability weights.

The data in Table 2.2 are consistent with the broad hypothesis in our study – patients who undergo a more invasive treatment for CAD are more likely to quit smoking. However, they could also be consistent with a story in which people who undergo CABG are also more likely to quit smoking for reasons unrelated to their surgery. If our hypothesis is true, we should see that the differential quitting behavior between CABG, PCI, and medically managed patients is driven by quits that occur close to the date of the treatment.
Table 2.3 – Quitting Within Six Months of Treatment

<table>
<thead>
<tr>
<th>Treatment</th>
<th>MM</th>
<th>PCI</th>
<th>CABG</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoke before treatment</td>
<td>15.7%</td>
<td>15.6%</td>
<td>15.1%</td>
<td>15.7%</td>
</tr>
<tr>
<td>Smoke after treatment</td>
<td>14.2%</td>
<td>12.9%</td>
<td>10.5%</td>
<td>13.8%</td>
</tr>
<tr>
<td>Percentage Point Change</td>
<td>-1.6%</td>
<td>-2.7%</td>
<td>-4.5%</td>
<td>-1.9%</td>
</tr>
<tr>
<td>Percent Change (Quit Rate)</td>
<td>10.1%</td>
<td>17.4%</td>
<td>30.1%</td>
<td>11.8%</td>
</tr>
<tr>
<td>Count</td>
<td>10,713</td>
<td>771</td>
<td>781</td>
<td>12,265</td>
</tr>
</tbody>
</table>

Note: This table includes every NHIS respondent who was diagnosed with CAD in our data prior to their interview date. It shows their smoking status exactly six months before and exactly six months after diagnosis (CAD) or surgery (PCI/CABG). Data are weighted by the NHIS probability weights.

Table 2.3 focuses on those quits that take place immediately around the initiation of treatment, where the initiation of treatment is defined to be the diagnosis date for medically managed patients and the procedure date for PCI and CABG patients. The “before” period is exactly six months before the treatment date, while the “after” period is exactly six months after treatment date. Among the 10,713 patients diagnosed with CAD who receive only medical management, approximately 15.7 percent smoked six months before their diagnosis and 14.2 percent smoked six months after their diagnosis. The quitters represent a 1.6 percentage point reduction in the number of smokers and a 10.1 percent reduction. The corresponding numbers for PCI are a 2.7 percentage point reduction and a 17.4 percent reduction. For CABG, they are a 4.5 percentage point reduction and a 30.1 percent reduction.

---

34 Creating a “quit window” around the treatment is necessary for two reasons. First, it is unlikely that many individuals quit on exactly the day their treatment began. Second, our smoking data, which are based on individuals’ recollections, are insufficiently precise to pinpoint the exact day of quitting. Our conclusions do not change with other reasonable definitions of the quit window.
Further evidence is provided by Figures 2.2 and 2.3. In Figure 2.2, we calculate the percentage of the population smoking at twelve points in time, measured in years relative to the date of diagnosis (in the case of medically managed patients) or procedure (in the case of PCI and CABG patients).\textsuperscript{35} In the CABG series, for example, the year -3.5 shows the percentage of CABG patients who were smoking exactly three and a half years prior to their procedure date. In the 10 years prior to the treatment date, the three series track each other reasonably closely. At

\textsuperscript{35}Because we have data on only the most recent quit date for each individual, we assume that each smoker was smoking in all years before their quit date. Since we are using Medicare data for our analysis, most people are over 65 when they received their diagnosis or procedure, and it is unlikely that they started smoking for the first time in the ten years immediately prior. It is possible that individuals quit and restarted during this time period, and we do not distinguish them from continuous smokers.
the first point on the graph – 9.5 years before treatment – CABG patients were roughly one percentage point more likely than PCI patients to smoke, who were, in turn, roughly one percentage point more likely than MM patients to smoke. In the period immediately prior to treatment, the three groups smoked at roughly equal rates. The differences between the three series emerge most starkly in the period immediately after treatment initiation. Six months after their treatment begins, CABG patients were substantially less likely to smoke than either PCI or MM patients.

Figure 2.3 – Quit Rate by Year Relative to Diagnosis (MM) or Procedure (PCI & CABG)
Figure 2.3 displays the same data in a different format, showing the annual quit rate for patients in each of the three groups relative to the date of treatment. For the group that received only medical management, roughly 5 percent of smokers quit each year in the nine years prior to being diagnosed, a rate that doubled to 10 percent during the year of their diagnosis with coronary artery disease. The PCI and CABG series show a similar trend, though they represent fewer individuals and are somewhat noisier. In the years prior to their procedure, roughly 5 percent of smokers quit each year, though this percentage began to rise the year before the procedure date. During the procedure year – defined to be the six-month window on either side of the procedure date – the quit rate jumped to 17 percent for patients receiving PCI and 30 percent for patients receiving CABG. In the year following treatment, the quit rate for all three groups dropped back to approximately 5 percent. Figures 2.2 and 2.3 provide reasonably compelling evidence that at least a portion of the increased quit rate for more invasive treatments observed in Table 2.2 is related to treatment received, and not simply a spurious correlation.

2.4 Results

To further explore the relationship between treatment for coronary artery disease and smoking behavior, we fit two related models of quitting smoking. The first is a discrete time linear probability hazard function with 11 periods: 9 before treatment and 2 after. We fit the model using individual data, allowing us to control for time-varying events – like CAD diagnosis or a patient’s first AMI – that may occur prior to, concurrent with, or after a PCI or CABG procedure. The second model is a multi-period quit function using grouped data, inspired by Donald and Lang (2007). In this model, we build a synthetic panel, grouping the individual data by period and treatment type into 33 cells (3 treatment types and 11 periods), and run difference-
in-differences regressions with 11 data points.

Before presenting the results of these models, it is useful to point out the relationship between a smoking participation function and a quit function. As an identity\(^{36}\)

\[
\frac{s_t}{s_{t-1}} \equiv 1 - q_t, \quad \quad \quad (2.1)
\]

where \(s_t\) and \(s_{t-1}\) are the smoking participation rates in periods \(t\) and \(t-1\), respectively, and \(q_t\) is the quit rate in the window defined by periods \(t\) and \(t-1\). All rates are defined as fractions and can be interpreted as probabilities at the individual level. Take natural logarithms of the identity to obtain

\[
\ln \frac{s_t}{s_{t-1}} \equiv \ln(1 - q_t) \cong -q_t. \quad \quad \quad (2.2)
\]

The approximation in the last part of Equation 2.2 is better for smaller values of \(q\) (\(q \leq 0.2\) is generally used as a rule of thumb). However, even for a quit rate as large as 0.3 (the largest rate in our data), \(\ln(1 - q) = -0.350\), which is close to 0.3. Equation 2.2 indicates that a regression in which the first difference of the log of smoking participation is the dependent variable should have approximately the same coefficients with the signs reversed as one in which

---

36 Let \(S_t\) be the number of smokers in period \(t\), and let \(S_{t-1}\) be the number of smokers in period \(t-1\). Let \(Q_t\) be the number of quitters in period \(t\) (the number who smoke in period \(t-1\) but do not smoke in period \(t\)). Assume as is the case in our data that there are no starters or re-starters. Then

\[S_t = S_{t-1} - Q_t.\]

Divide both sides of the identity by \(N\), the size of the population:

\[
\frac{S_t}{N} = \frac{S_{t-1}}{N} - \frac{Q_t}{N} \frac{S_{t-1}}{N}.
\]

Solve the last identity for \(\frac{N}{S_t} \frac{N}{S_{t-1}}\) to obtain equation (1).
the quit rate is the dependent variable. It also suggests that it is useful to begin with a log
smoking participation function to arrive at a specification of a quit function. In particular, if the
log smoking participation function contains individual fixed effects, these effects are eliminated
by taking first differences to obtain the quit function.

2.4.1 Discrete-time Linear Probability Hazard Model (Individual Data)

To implement the first model, we develop a synthetic panel with 12 points in time, as
shown in Figure 2.2. For CABG and PCI patients, there are 10 points in time prior to treatment
(from 9.5 years before to 0.5 years before) and two points in time after treatment (0.5 years after
and 1.5 years after). For medically managed (MM) patients, there are the same 10 points in time
before diagnosis and the same two points in time after diagnosis. To focus on the key aspects of
the model, we ignore the socioeconomic and demographic variables for the time being, assume a
single Elixhauser comorbidity, and suppress the subscript i for an individual. Let $f_g$ be a person-
specific fixed effect. Let $a_t$ be a dummy variable that equals 0 at each point in time before
treatment for PCI or CABG and equals 1 at each point in time after treatment. Specifically, $a_t$
equals 1 in periods 11 and 12. This variable is not relevant for MM patients (see below). Let $d_t$
be a dummy variable that equals 0 at all points in time before diagnosis and equals 1 at all points
in time after diagnosis. Let $h_t$ be a dummy variable that equals zero before a patient’s first AMI
(in our data) and equals 1 after a patient’s first AMI. Finally, let $e_t$ equal 0 before an Elixhauser
comorbidity is reported and equals 1 thereafter.

The log smoking participation model for PCI and CABG patients ($g = p$ or $c$)

$$\ln s_{gt} = f_g - \beta_{gt} a_{gt} - \phi h_{gt} - \gamma e_{gt} - \alpha d_{gt} - \lambda t,$$ (2.3)
where we assume a linear trend in the absence of treatment. The model for MM patients is the same except that $a_{mt}$ coincides with $d_{mt}$, so that we constrain $\beta_m$ to equal zero. After pooling and taking first differences, one obtains

$$-(\ln s_t - \ln s_{t-1}) \equiv q_t = \lambda + \alpha(d_t - d_{t-1}) + \beta_p p(a_t - a_{t-1}) + \beta_c c(a_t - a_{t-1}) + \phi(h_t - h_{t-1}) + \gamma(e_t - e_{t-1}).$$ (2.4)

Strictly speaking, time-invariant individual characteristics, such as formal schooling, can only be added to Equation 2.4 by assuming that they interact with the linear trend in Equation 2.3. Our results are not affected by allowing the trend to be nonlinear or by allowing individual characteristics to interact with the indicator for the period after treatment in addition to their interactions with a linear trend.

We fit Equation 2.4 as a discrete time linear probability hazard model. We include only individuals who smoke at the first point in time for which we compute smoking participation (9.5 years before treatment or diagnosis), dropping everyone who never smoked or had previously quit. Each person is assigned a $q_{it}$ variable that is equal to one in the period in which they quit and zero in all other periods. Individuals are deleted once they quit. The model in Equation 2.4 has at most 11 observations per person corresponding to the 11 time periods in Figure 2.3. Individuals who smoke in all periods are the censored observations. The first period is defined by the window starting 9.5 years before treatment and ending 8.5 years before treatment. The last period is the window from 0.5 years after treatment to 1.5 years after treatment. The key window is period 10 and spans the dates from half a year before treatment or diagnosis to half a year after. That is the only period in which $a_t - a_{t-1}$ is equal to 1. Since there are repeat
observations on all persons except those who quit in period 1, we cluster standard errors at the individual level. Standard errors that ignore clustering are, however, very similar to those that take account of it. This would be the case if the unspecified disturbance term in the log smoking participation function in Equation 2.3 is a random walk. In that case, we eliminate serial correlation by taking first differences.

Results are shown in Table 2.4. Note that $\Delta$After stands for $a_t - a_{t-1}$, $\Delta$Diagnosed stands for $d_t - d_{t-1}$, and $\Delta$AMI stands for $h_t - h_{t-1}$ in the table. Column 1 shows the simplest specification, with controls only for diagnosis with CAD and treatment with either PCI or CABG. The $\Delta$Diagnosed indicator is equal to one in the period when MM patients, PCI patients, or CABG patients are diagnosed, whether the diagnosis occurs in the same period as the treatment or not.\(^{37}\) Being diagnosed with CAD is associated with a 5.0 percentage point increase in the probability of quitting, on top of the typical yearly quit rate of 4.9 percent (reflected in the coefficient on the constant term). Being treated with PCI or CABG is associated with an incremental 9.0 or 22.5 percentage point increase in the quit rate, respectively. This assumes that diagnosis and treatment occur in the same period. If not, the diagnosis coefficient must be subtracted since $\Delta$Diagnosed equals zero for PCI and CABG patients diagnosed in a period prior to their treatment period but equals one for all MM patients. That results in a 4.0 percentage point increase for PCI patients and a 17.5 percentage point increase for CABG patients. Since 40 percent of PCI patients and 43 percent of CABG patients are diagnosed before treatment, the average percentage point increases in the quit rates are 7.0 and 20.4, respectively.\(^{38}\)

\(^{37}\) In results not shown, we find no evidence of a differential diagnosis effect based on treatment received. Therefore, in all results shown, we assume that there is a single diagnosis effect that does not vary by treatment.

\(^{38}\) For MM patients, the increase in the quit rate is 0.0501. Since 60 percent of PCI patients are diagnosed in the same period as treatment, the average predicted increase in the quit rate for these patients is $0.0897 + 0.60 \times 0.0501 = 0.1198$. The difference between that increase and the increase for MM patients is 0.0697 or 7.0 percentage points. Since 57 percent of CABG patients are diagnosed in the same period as treatment, the average predicted increase is
of how these computations are made, the difference between the coefficients on PCI and CABG treatments is 13.5 percentage points and is significant at the $\alpha = 0.05$ level.

Table 2.4 – Regression Results for Discrete Time Linear Probability Hazard Model

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.049***</td>
<td>0.049***</td>
<td>0.047***</td>
<td>0.046***</td>
<td>-0.050</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td>(0.001)</td>
<td>(0.001)</td>
<td>(0.011)</td>
<td>(0.032)</td>
</tr>
<tr>
<td>PCI * ∆After</td>
<td>0.090**</td>
<td>0.054</td>
<td>0.034</td>
<td>0.033</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>(0.038)</td>
<td>(0.040)</td>
<td>(0.040)</td>
<td>(0.039)</td>
<td>(0.042)</td>
</tr>
<tr>
<td>CABG * ∆After</td>
<td>0.225***</td>
<td>0.195***</td>
<td>0.160***</td>
<td>0.155***</td>
<td>0.141***</td>
</tr>
<tr>
<td></td>
<td>(0.045)</td>
<td>(0.045)</td>
<td>(0.044)</td>
<td>(0.044)</td>
<td>(0.046)</td>
</tr>
<tr>
<td>∆Diagnosed</td>
<td>0.050***</td>
<td>0.042***</td>
<td>0.025***</td>
<td>0.025***</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>(0.008)</td>
<td>(0.008)</td>
<td>(0.008)</td>
<td>(0.026)</td>
</tr>
<tr>
<td>∆AMI</td>
<td>0.083***</td>
<td>0.075***</td>
<td>0.076***</td>
<td>0.076***</td>
<td>0.076***</td>
</tr>
<tr>
<td></td>
<td>(0.023)</td>
<td>(0.024)</td>
<td>(0.024)</td>
<td>(0.024)</td>
<td>(0.024)</td>
</tr>
<tr>
<td>CABG * ∆After - PCI * ∆After</td>
<td>0.135**</td>
<td>0.141**</td>
<td>0.126**</td>
<td>0.122**</td>
<td>0.115**</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elixhauser</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Demographics</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Year, Period, Age, Yrs Smoked</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Observations</td>
<td>26,658</td>
<td>26,658</td>
<td>26,658</td>
<td>26,658</td>
<td>26,658</td>
</tr>
<tr>
<td>Individuals</td>
<td>3,065</td>
<td>3,065</td>
<td>3,065</td>
<td>3,065</td>
<td>3,065</td>
</tr>
</tbody>
</table>

Robust standard errors, clustered at the individual level, in parentheses. Regressions are weighted by NHIS probability weights. ∆AMI is an indicator for a patient having her first AMI (in our data) in a particular period. Specifications 3 and 4 include 29 dummy variables indicating when the Elixhauser comorbidity conditions were first diagnosed (in our data). In specification 4, demographic controls include gender, race, education dummies, and income category dummies (including a dummy for missing income data). In specification 5, we also include year dummies, period dummies, age and age squared, and a variable indicating the number of years a person had been smoking as of the baseline period (9.5 years before treatment). *CABG * ∆After - PCI * ∆After* is the difference between the PCI * ∆After and CABG * ∆After coefficients. * significant at the 0.10 level; ** significant at the 0.05 level; *** significant at the 0.01 level.

In columns 2-5 of Table 2.4, we add additional control variables that may be associated with both selection into treatment and the propensity to quit smoking. In column 2, we add a control for a patient’s first AMI. In column 3, we further control for the Elixhauser comorbidity

\[0.2250 + 0.57 \times 0.0501 = 0.2536.\] The difference between that increase and the increase for MM patients is 0.2035 or 20.4 percentage points.
conditions. In column 4, we add controls for demographic characteristics. And, in column 5, we add controls for the period, the calendar year, an individual’s age and age squared as of the treatment date, and the number of years she has smoked as of the baseline period (9.5 years prior to treatment). An AMI proves to be a strong predictor of quitting smoking, increasing the predicted probability of quitting by between 7.5 and 8.3 percentage points, depending on the specification. Moreover, once we add AMI, the estimated size of the PCI coefficient declines and is no longer statistically significant at conventional levels. In our final specification, we find that patients undergoing PCI are 2.6 percentage points more likely to quit smoking, though given the imprecision of our estimate, we cannot rule out the possibility of no association. While the coefficient on CABG also declines, it remains large and strongly statistically significant, both in comparison to medically managed patients and PCI patients. In our fifth specification, undergoing CABG surgery in the same period in which CAD was diagnosed is associated with a roughly 14.1 percentage point increase in the probability of quitting smoking. The increased quit rate associated with CABG is substantially larger than the increase associated with less invasive treatments for CAD, and approximately twice the increase associated with an AMI. Moreover, in every specification, the difference between the CABG and PCI coefficients is consistently large – ranging from 0.115 to 0.141 – and significant at the 0.05 level.

2.4.2 Multi-Period Quit Model Using Grouped Data

To illustrate that our results are not sensitive to a flexible specification of period effects and to account for clustering of disturbance terms by group and period at the individual level, we aggregate the data into 11 periods for each of the three groups of patients. There are 33 cells in the aggregate sample, corresponding to the 33 data points shown in Figure 2.3. We obtain two

39 As discussed above, this estimate is for a PCI patient who was diagnosed and treated in the same period.
quit series. The first is unadjusted for covariates and is identical to the quit rates that appear in Figure 2.3. The second quit series adjusts for effects due to diagnosis, AMI, and Elixhauser comorbidities. It is obtained from the individual data by estimating a discrete time hazard function for the probability of quitting that includes 11 period dummies interacted with each of three treatment dummies (one for CABG, one for PCI, and one for MM), and the diagnosis, AMI, and Elixhauser variables defined in Equation 2.4. The 33 coefficients associated with the period-treatment interactions are quit rates by group and period adjusted for the effects of the last three variables just mentioned.  

In the spirit of Donald and Lang (2007), we use this data to perform simple difference-in-differences regressions with 11 observations. To illustrate the model that we estimate, consider a log smoking participation function for two groups (g = c or p, c = CABG, p = PCI):

\[ \ln s_{gt} = \mu + \rho c + \beta c a_t + 11 \text{ period dummies} + \varepsilon_{gt}. \]  

(2.5)

Here \( a_t \), as defined in Equation 2.3, is an indicator that equals 1 in each of the two periods after treatment and \( \varepsilon_{gt} \) is the error term. Take the difference between each group in a given period to eliminate the intercept (\( \mu \)) and the period dummies. Then take first differences to eliminate the group effect (\( \rho \)):

\[ \ln s_{ct} - \ln s_{ct-1} - (\ln s_{pt} - \ln s_{pt-1}) \equiv q_{ct} - q_{pt} = \beta (a_t - a_{t-1}) + \text{error}. \]  

(2.6)

Equation 2.6 is a regression forced through the origin with 11 observations. The

---

40 We do not adjust for demographic and socioeconomic characteristics since the inclusion of these characteristics has a very minor impact on the adjusted quit rates.
dependent variable is the difference between the quit rate of CABG patients and the quit rate of PCI patients in each period. The independent variable, \((a_t - a_{t-1})\), equals 1 in the window spanning the period from 6 months before treatment to six months after treatment (period 10) and equals 0 in each of the other 10 periods or quit windows.

This approach has a number of attractive features. First, aggregation accounts for clustering in the disturbance term in an individual-level log smoking participation or quit equation by group and period. Second, if the error term in Equation 2.5 is a random walk, then serial correlation is eliminated once first differences are taken. Third, the regression specified by Equation 2.6 implicitly controls for a full set of period effects. Finally, by focusing on the difference in the quit rates in each period, we are asking whether this difference during the treatment period is sufficiently unusual compared to past and future values that it is unlikely to have arisen by chance. If the quit rates in the two series normally track one another but do not during the treatment year, we would expect that there is something unusual about the treatment year. On the other hand, if the quit rates in the two series often diverge wildly, then a substantial divergence in the treatment year might simply be due to chance.

Six aggregate quit regressions are contained in Table 2.5. The three in the top row use the unadjusted quit series, while the three in the bottom row use the adjusted quit series. In column 1, the dependent variable is the difference between the CABG and PCI quit rates; in column 2, it is the difference between the CABG and MM rates; and in column 3, it is the difference between the PCI and MM rates. Three separate regressions are obtained for each series because of evidence that the residual variance is not the same for each dependent variable.\(^{41}\) To be consistent with the notation in Table 2.4, the variable \(a_t - a_{t-1}\) is termed \(\Delta\)After

\[^{41}\text{Consider the following two regressions}
q_{pt} - q_{p,t-1} - (q_m - q_{m,t-1}) = \beta_p(a_t - a_{t-1})\]
in the table.

Table 2.5 – Quit Rate Regression with Grouped Data

<table>
<thead>
<tr>
<th></th>
<th>(1) CABG - PCI</th>
<th>(2) CABG - MM</th>
<th>(3) PCI - MM</th>
</tr>
</thead>
<tbody>
<tr>
<td>∆After (no adjustments)</td>
<td>0.127**</td>
<td>0.200***</td>
<td>0.073</td>
</tr>
<tr>
<td></td>
<td>(0.050)</td>
<td>(0.044)</td>
<td>(0.047)</td>
</tr>
<tr>
<td>∆After (with adjustments)</td>
<td>0.120**</td>
<td>0.142***</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>(0.052)</td>
<td>(0.042)</td>
<td>(0.051)</td>
</tr>
<tr>
<td>N</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
</tbody>
</table>

Note: Each cell represents the coefficient on ∆After from a separate regression, with the standard error in parentheses. The dependent variable is the quit rate for one group of patients minus the quit rate for another. The independent variable is a dummy for the treatment year (ΔAfter). In the top row, the quit rate is unadjusted. In the bottom row, the quit rate is adjusted for diagnosis, AMI, and Elixhauser comorbidities, based on an individual regression. All quit rates are calculated using the NHIS probability weights. Regressions are forced through the origin and weighted by the inverse of the square root of the variance of the difference between period-specific quit rates from individual level regressions with 33 group period interactions. * significant at the 0.10 level; ** significant at the 0.05 level; *** significant at the 0.01 level.

Focusing on the first column of Table 2.5, one sees that the difference between the CABG quit rate and the PCI quit rate in treatment period is between 12 and 13 percentage points, regardless of whether one uses the raw data or adjusts the quit rate for covariates. In the second column, we see that the difference in the quit rate between CABG and MM patients drops when one adjusts for covariates, but remains large (14.2 percentage points) and strongly statistically significant. In the final column, we see that while the unadjusted quit rate for PCI patients is 7.3

\[ q_{ct} - q_{ct-1} - (q_{mt} - q_{mt-1}) = \beta_c(a_t - a_{t-1}), \]

where m denotes medical management. Estimates of \( \beta_p, \beta_c, \text{ and } \beta_c - \beta_p \) could be obtained from a pooled regression of the form

\[ q_{ct} - q_{ct-1} - (q_{mt} - q_{mt-1}) = \beta_c(a_t - a_{t-1}) + \beta_p(1-c)(a_t - a_{t-1}). \]

We do not follow that approach because the residual variance in the first regression is not equal to the corresponding variance in the second regression.
percentage points larger than for MM patients, once one adjusts for a patient’s CAD diagnosis and other medical conditions, the difference between the two quit rates is substantially smaller and not statistically distinguishable from zero.

2.5 Discussion

2.5.1 Selection Bias

When interpreting our results, one important issue to address is the potential for selection bias. CABG patients are clearly different from PCI patients on a number of observable characteristics, as shown in Table 2.1. While we control for these variables in our regressions, are there unobservable characteristics that lead CABG patients to be more likely to quit smoking for reasons unrelated to their surgery? In particular, are CAD patients – or their doctors – selecting CABG over PCI in instances where the patient is more likely to quit smoking?

Patient selection bias is unlikely to explain our results for several reasons. First, as shown in Figure 2.2, the smoking participation rates for CABG and PCI patients are very similar in the period immediately prior to their procedure. Since both groups having roughly 15% of their population still smoking six months prior to their procedure, there is nothing about the smoking participation rate that, in and of itself, implies a difference in the probability to quit. Second, as shown in Figure 2.3, the annual quit rates are very similar between CABG and PCI patients going back nine periods prior to their procedure. While the quit rate for the CABG group rises between periods 7 and 9, it remains fairly close to the PCI quit rate and the difference is never statistically significant.\(^{42}\) Finally, CAD patients have limited authority to select their procedure, since the recommendation is made by their doctor. While a patient can certainly

\(^{42}\) To test the statistical significance of the difference, we run a series of regressions like those in Table 4, that include the interaction of an indicator variable for Periods 7, 8, and 9 with an indicator for CABG and PCI. The results are shown in Table 6.
refuse any treatment she is referred to, this would most likely happen in the case where a patient was referred to CABG and refused due to the short-term health risks. However, it seems likely that those patients who are concerned about the short-term health risks of CABG are also those who are most likely to quit smoking, a bias that would work against our main finding.

Doctor selection bias is also unlikely to explain our results. First, doctors who refer patients to PCI or CABG generally do so on the basis of the severity of CAD, not based on a patient’s comorbidities or health behaviors. When we enter separate CAD diagnosis effects for PCI and CABG patients into the regressions in Table 2.4 – to capture whether the severity of the diagnosis impacts smoking behavior – we do not find any significant difference. Second, doctors who perform CABG (cardiothoracic surgeons) and PCI (interventional cardiologists) could selectively refuse those patients who they think are unlikely to quit smoking. Putting aside the difficulty in accurately forecasting who will quit smoking, one would expect selective refusal to occur more frequently where there is excess demand. Between 2001 and 2008, Epstein et al. (2011) estimate that the rate of PCI procedures in the U.S. held steady while the rate of CABG procedures decreased by 33%. Given these trends, interventional cardiologists (who perform PCI) are more likely to face excess demand than cardiothoracic surgeons (who perform CABG) and therefore more likely to engage in patient selection on the basis of factors – like propensity to quit smoking – that would be likely to improve a patient’s outcome. This bias, should it be present, would tend to work against our main finding.

2.5.2 Alternative Mechanisms

In this paper, we demonstrate that CABG patients are substantially more likely to quit smoking than those who have less invasive treatment for CAD. When combined with other
potential lifestyle modifications, this change in behavior may help explain why CABG patients experience better mortality and morbidity outcomes over the long term than otherwise similar PCI patients. This finding is consistent with a model in which patients undergoing CABG receive a stronger and more persistent signal about the severity of their disease, leading them to improve their behavior. However, other explanations are also possible.

One alternative explanation is that CABG patients quit because they are forced to go “cold turkey” during the length of their hospital stay. For the week that they are in a hospital, most CABG patients are presumably unable to smoke, potentially kick-starting the quitting process and making it easier for them to stay off cigarettes when discharged. PCI patients, by contrast, typically spend a day or less in the hospital, which might account for their lower quit rate in the period surrounding their procedure.

To test this theory – and provide a falsification test for our main results – we study two common procedures that are unrelated to smoking but require relatively long hospital stays: total knee replacement and total hip replacement. These procedures – taken together – were performed over one million times in 2010, with over 500,000 procedures taking place in patients 65 and older (CDC, 2010). These two surgeries – which involve replacing a knee or hip joint with a metal or plastic replica – are generally done to relieve arthritic pain, and patients spend at least three to four nights in the hospital recovering. Smoking is largely unrelated to knee or hip pain and so undergoing knee or hip replacement surgery would be unlikely to convey any new information to a smoker about the value of quitting.

43 We use ICD-9-CM procedure code 81.51 for Total Hip Replacement and 81.54 for Total Knee Replacement
44 Thanks to Isaac Sorkin, who suggested we find non-smoking related procedures to do this type of analysis.
To test the impact of a long hospital stay on smoking, we identify total knee and total hip replacement patients using the same linked NHIS / Medicare data as in the main analysis. Figure 2.4 reproduces the CABG quit rate from Figure 2.3 and adds a series showing the quit rate for patients who undergo knee or hip replacements.\textsuperscript{46} When we combine knee and hip replacement recipients, we have data on 155 patients who smoked nine and half years prior to their surgery, slightly below the 206 CABG smokers in our data. In the years prior to their procedure, the quit rate for knee/hip and CABG patients are fairly similar, with the knee/hip quit rate somewhat

---

\textsuperscript{46} The small number of patients who have both knee and hip replacements are assigned a procedure date based on their total knee replacement.
higher five to nine years prior to surgery and the CABG quit rate somewhat higher one to three years prior to surgery. However, while the quit rate for CABG patients jumps dramatically in the one year window surrounding their procedure, we see no similar increase for knee/hip patients. Consistent with the visual evidence shown in Figure 2.4, regression results (not shown) show no significant relationship between total knee or hip replacement surgery and quitting. It appears unlikely that a substantial portion of the observed CABG impact on quitting is caused by going “cold turkey” during a patient’s hospital stay.

Alternatively, one might explain our main results by the doctor a patient sees. As discussed above, CABG procedures are performed by cardiothoracic surgeons, while PCI procedures are performed by interventional cardiologists. If these two types of doctors systematically give different advice to smokers on the importance of quitting, this might be the mechanism through which the CABG quit rate rises above the PCI quit rate near the treatment date. Though different than our main hypothesis – that CABG sends a stronger and more persistent signal to the patient about his or her health problems – this mechanism is similar. Rather than the scar on a CABG patient’s chest, it may be her heart surgeon who convinces her to quit smoking. Either way, the patient updates her opinion on the value of quitting smoking based on the CABG experience. And, while this alternative explanation – that heart surgeons are more persuasive smoking cessation advocates – seems less likely to us, it arguably provides more actionable policy recommendations.

2.5.3 Quitting and the PCI-CABG Mortality Differential

According to Figure 2.1, the cumulative mortality rate of PCI patients five years after surgery is approximately 6 percentage points higher than that of CABG patients (26 percent
versus 20 percent). How much of this differential might be due to quitting behavior? Let \( m \) be the mortality rate after five years, \( m_s \) be the mortality rate of those who smoke one period after surgery, \( m_q \) be the mortality rate of quitters, \( m_n \) be the mortality rate of former smokers and nonsmokers, \( s \) be the fraction who smoke one period before surgery, and \( q \) be the fraction of smokers who quit. As an identity

\[
m = s m_s + s q (m_q - m_s) + (1 - s) m_n.
\] (2.7)

Assume that the only difference between the mortality rate of PCI patients (\( m_p \)) and the mortality rate of CABG patients (\( m_c \)) is due to differential quit behavior (\( q_c > q_p \)). Then

\[
m_p - m_c = s (q_p - q_c) (m_q - m_s).
\] (2.8)

In our data, approximately 15 percent of CABG and PCI patients smoke one period before surgery (\( s = 0.15 \)). Our estimates indicate that \( q_p - q_c = -0.12 \). The medical literature suggests that CAD patients who quit smoking following surgery have a cumulative five-year mortality rate that is 20 percentage points smaller than patients who continue to smoke: \( m_q - m_s = -0.20 \) (Chen et al., 2012; Critchley and Capewell, 2003; de Boer et al., 2013; van Domburg et al., 2000). Hence, \( m_p - m_c \) equals 0.36 percentage points. Put differently, the higher quit rate of CABG patients compared to PCI patients accounts for roughly 6 percent of the 6.0 percentage point differential in mortality between the two groups of patients.

According to the computation just made, the higher quit rate of CABG patients makes a minor contribution to their improved survival prospects relative to PCI patients. Several factors should be kept in mind in an evaluation of this result. First, the component of the mortality differential due to quit behavior is small because the smoking participation rate is small and would be much larger in populations with higher rates. For example, in one evaluation of the impacts of quitting smoking on survival among PCI patients in China, approximately 60 percent
of males smoked before the intervention (Chen at al., 2012). Second, quitting after surgery has bigger effects on survival than the use after surgery of statins, beta-blockers, or enzyme inhibitors (for example, Critchley and Capewell, 2003). Finally, other lifestyle changes may accompany quitting smoking.

2.6 Conclusion

Coronary Artery Disease is a frequently occurring and deadly disease. There are several common treatments – including medical management, PCI, and CABG – and each has benefits and costs associated with it. In this paper, we have examined one previously unexplored benefit of more invasive treatment: those who undergo a procedure, particularly the more invasive CABG surgery, are more likely to quit smoking. In our preferred regression model, we estimate that CAD patients who undergo CABG are 12 percentage points more likely to quit smoking in the one-year window surrounding their surgery than patients who undergo PCI. During the same one-year window, CABG patients are 14 percentage points more likely to quit smoking than medically managed patients. These results are robust to a number of alternative specifications.

While we do not have data on behaviors other than smoking, we suspect that patients undergoing more invasive surgery are also more likely to improve their diet, limit excessive consumption of alcohol, and (when recommended) exercise more. Taken together, these behavioral responses may offset the inherent risks of a more invasive surgery and help explain why the longer term outcomes for CABG patients exceed those of similar patients receiving PCI. Our findings also highlight the importance of emphasizing healthier behaviors to those patients who have less invasive medical treatment.
3. Schools and Obesity: A Natural Experiment Using the New York City High School Admissions Process

Jesse Margolis
3.1 Introduction

Obesity in the United States has been on the rise in recent decades, with the percentage of adults over 20 classified as obese (Body Mass Index – BMI – over 30) increasing from 22% in 1988-1994 to 34% in 2007-2008. Over the same period, the percentage of obese adolescents, age 12-19 years old, increased from 11% to 18%. Obesity has been linked to a greater risk of health problems, including increased rates of heart disease and type 2 diabetes (Center for Disease Control and Prevention: CDC, 2010). The CDC has written that “changes in children’s physical activity and eating habits over time appear to contribute to increases in prevalence of obesity” (CDC, 2010). Since children and adolescents spend six to nine hours a day in and around their school, one might expect that their school environment has an impact on their exercise patterns and caloric intake.

Recent research has found that the characteristics of a child’s school affect her weight and probability of being obese. Using a large dataset from California, Currie et al. (2010) find that if a fast food restaurant is located within 0.1 miles of a high school, the proportion of 9th graders who are obese increases by 5.2 percent. Anderson and Butcher (2006) find that a 10 percentage point increase in the probability that a student is exposed to junk food in her school increases her BMI by one point, on average. Schanzenbach (2009) finds that students who consume school lunches – rather than bringing their own lunch from home – weigh more and are more likely to be obese. Drake et. al. (2012) estimate that obesity rates would decline by 26% if adolescents played on at least two sports teams per year. In each case, the specific characteristics of a school – its location, junk food policies, school lunch quality, and athletic opportunities – appear to affect student weight, BMI, and/or obesity (Anderson et al., 2011).

47 I would like to thank Michael Grossman, David Jaeger, Ted Joyce, Wim Vijverberg, Steve O’Connell, Shannon O’Grady and participants at the 2015 American Educational Research Association Annual Meeting for helpful comments on an earlier version of this paper. Any errors are my own.
Figure 3.1 – Distribution of PSAL Sports teams by school and incoming obesity rates by program in New York City

Note: this chart only includes schools and programs that had lotteries included in this study. PSAL sports team data are from the 2014/15 high school directory. Incoming obesity rates are from 2011 NYC FITNESSGRAM.

Based, in part, on the belief that schools influence obesity, the New York City Department of Education (NYCDOE) has developed and expanded system-wide programs to help children develop healthier lifestyles. The city has introduced healthier school lunches, promoted physical activity, and provided fitness reports to all students and parents using the NYC FITNESSGRAM program. There are also a number of school-level channels in New York through which attending one school or another might affect a student’s weight. Certain schools are located in close proximity to fast food restaurants, while others are more distant. The availability of sports programs varies widely by school, as shown on the left side of Figure 3.1. Also in Figure 3.1, we can see that the body composition of each school’s incoming 9th grade class also varies substantially, leading to potential peer effects on student weight (Christakis and
Furthermore, individual schools have undertaken their own initiatives to reduce obesity and improve student health.\textsuperscript{49}

The goal of this research is to begin to answer the question: do schools affect obesity within New York City? In particular, I ask whether attending one New York City high school versus another has a differential effect on a student’s weight, BMI, and probability of becoming overweight or obese. To answer this question, I use a natural experiment provided by the admissions process to New York City high schools. For a subset of high schools that admit students by lottery, I use a t-test to estimate whether there is a statistically significant impact of attending that high school on a student’s weight, BMI, overweight-ness, and obesity. I find that one observes statistically significant effects at the 0.05 level roughly as often as one would expect if schools had no differential effect on obesity-related variables (i.e. about 5\% of the time). This finding is consistent with the proposition that the specific school a student attends does not have a substantial effect on his or her weight gain. For comparison, I perform a similar analysis for student achievement variables – including credits earned, test scores, attendance, and on-time graduation – where \textit{a priori} one would expect a differentiated school effect. For achievement variables – particularly credits earned in 9\textsuperscript{th} grade – I find a statistically significant effect more often than one would expect by chance.

\textbf{3.2 Methodology}

In this study, I take advantage of the lottery component of the high school admissions

\begin{footnotesize}
\begin{itemize}
\item[48] See Cohen-Cole and Fletcher (2008) for an argument against peer effects on student weight.
\item[49] For example, the \textit{New York Times} profiled the non-profit FoodFight, which has been hired by several New York City schools, saying that their curriculum “blends media literacy, politics, nutrition and cooking” to encourage more thoughtful, healthy eating (October 26, 2011). Another non-profit, New York Road Runners, works with 450 New York City schools to promote exercise (\textit{New York Times}, October 22, 2011).
\end{itemize}
\end{footnotesize}
process in New York City. Each year, approximately 80,000 eighth graders rank up to 12 high school programs in priority order. Using a computer algorithm based on the national medical residency match, the New York City Department of Education (NYCDOE) then assigns each applicant to a high school (Abdulkadiroğlu, Pathak, and Roth, 2004). When certain programs are oversubscribed – a list that includes a large number of new small schools started since 2003 – admission is based partly on a random number assigned to each student.

For those schools that admit students by lottery and are oversubscribed, I use the admissions process as a natural experiment. For each school, I divide students who chose the school first into three groups: 1) students who were subject to the lottery and were admitted, 2) students who were subject to the lottery and were not admitted, and 3) students who were not subject to the lottery (whether or not they were admitted). For each school, the first group forms the “treatment” group and the second group forms the “control” group. Students in the third group are not part of this study, since their admission was not based on a random number. This process is described in detail in Section 3.4.

I then compare the outcomes of those two groups on five health related measures: student height, weight, BMI, an indicator for being overweight, and an indicator for obesity. I also compare each group along a series of academic outcomes in high school, including attendance, credits earned, English test scores, and graduation rates. For each school in each year where students were admitted through an oversubscribed lottery, I perform a t-test on the null hypothesis of equal means, assuming equal variances (Kenkel, 1981: 412). The test statistic is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)S^2_{X_1}}{n_1} + \frac{(n_2 - 1)S^2_{X_2}}{n_2}} \cdot \frac{1}{\sqrt{n_1 + n_2 - 2}} \cdot \sqrt{n_1 + n_2}}$$

(1.1)
In Equation 1.1, \( \bar{X}_1 \) is the average outcome of interest (weight, height, BMI, or an overweight or obesity indicator) for the treatment group and \( \bar{X}_2 \) is the same outcome for the control group. \( S^2_{\bar{X}_1} \) is the variance of the outcome for the treatment group and \( n_1 \) is the sample size, while \( S^2_{\bar{X}_2} \) and \( n_2 \) are the variance and sample size for the control group, respectively. I then obtain p-values by assuming \( t \) follows a Student’s t-distribution with \( n_1 + n_2 - 2 \) degrees of freedom.

3.3 Data

For this study, I was provided with de-identified student-level administrative data by the New York City Department of Education (NYCDOE). The data contain scrambled student ID numbers so that the data can be linked across years and across datasets.

To create the treatment and control groups, I use data from the high school admissions process for the years 2005 to 2009. These data contain one record for every 8\(^{th}\) grade applicant in the high school admissions process for each year. Each record has information on up to 12 choices made by the student in order of preference. For each choice, the data show the student’s eligibility, geographic priority, and the rank the school gave the student, where applicable. The file also shows the program the student was eventually matched to.

To calculate the outcome variables, I begin with student height and weight collected as part of the New York City FITNESSGRAM program. Through the program, which began in 2006, all students in Kindergarten through 12\(^{th}\) grade in New York City are supposed to have their height, weight, strength, endurance, and flexibility measured every year. Since it was first

---

50 The 2005 HSAPS process refers to the admissions cycle for students who will be enrolling in high school in September, 2005
introduced, participation has increased substantially, reaching over 80% in the 2009/10 and 2010/11 school years (see Table 3.1). I then calculate the Body Mass Index (BMI) for each student, defined as weight (in kilograms) divided by height squared (in meters). Using the BMI-for-Age growth charts provided by the Center for Disease Control and Prevention (CDC), I flag each student as being overweight if she is in or above the 85th percentile and obese if she is in or above the 95th percentile. For a student who is over 19 years old, I categorize her as overweight if her BMI is greater than or equal to 25 and obese if her BMI is greater than or equal to 30.

Table 3.1 – FITNESSGRAM Participation by Year

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>FITNESSGRAM Tests</td>
<td>453,635</td>
<td>632,168</td>
<td>743,307</td>
<td>830,572</td>
<td>811,764</td>
</tr>
<tr>
<td>K-12 Enrollment</td>
<td>994,292</td>
<td>981,862</td>
<td>975,421</td>
<td>983,664</td>
<td>986,821</td>
</tr>
<tr>
<td>Participation Rate</td>
<td>46%</td>
<td>64%</td>
<td>76%</td>
<td>84%</td>
<td>82%</td>
</tr>
</tbody>
</table>

Source: FITNESSGRAM tests based on the number of observations with a non-null value for the weight measure. K-12 Enrollment from the S-Form on the NYCDOE web site as of 1/22/2013 (excluding Pre-K students)

3.4 Confirming Randomization in School Lotteries

In New York City, students don’t technically apply to high schools, but rather to high school programs. Some high schools have more than one program, each of which may have a different academic focus and selection criteria. The new small high schools started since 2003 – the focus of this study – generally consist of a single program with an admissions criteria known as “limited unscreened.” For this group of schools, studying the impact of a program is identical to studying the impact of a school.

If a limited unscreened program is oversubscribed, student admission is determined first

by a student’s priority group and second by a random number assigned to each student. A student’s priority group is defined along two dimensions: geography and school rank. Certain programs give geographic priority to students who live in the school’s local district or borough. Moreover, schools with limited unscreened programs are allowed to rank their applicants with a 1 if they attended an information session or visited the school (known as making an “informed choice” to apply) and with a 2 otherwise. All else equal, students with higher geographic priority are admitted first over students with lower geographic priority and students with a rank of 1 are admitted before students with a rank of 2. In the 2005 admissions process, geographic priority was considered before school rank (i.e. a student with geographic priority 1 and rank 2 was admitted before a student with geographic priority 2 and rank 1). In the 2006 through 2009 admissions processes, school rank was considered before geographic priority.52

For limited unscreened programs that are under-subscribed – with fewer applicants than available seats – there is no lottery. All students are admitted to the program if they do not get into one of their higher ranked programs. For over-subscribed programs, the lottery occurs in a single priority group. All students in a higher priority group than the lottery group are admitted (if they do not match to a higher ranked program) and all students in a lower priority group are not admitted. For simplicity, in this study I focus only on students who selected a program as their first choice. Each student, therefore, can enter at most one lottery.

Table 3.2 shows the number of limited unscreened programs each year that were sufficiently over-subscribed to require that a lottery be held among their first-choice applicants. From 2005 to 2007, there were between 63 and 76 programs with first-choice lotteries, a number

52 In 2008, the NYCDOE added a third determinant of a student’s priority: where the student ranked the school on her list of 12 choices. Therefore, beginning in 2008, a student’s priority was determined by 1) the rank the school gave her, 2) where she lived, and then 3) where she ranked the school. Lotteries were only performed among groups that tied on all three priority levels. For this study, the practical impact is to limit the number of lotteries that occur among first choice applicants beginning in 2008 (see Table 2).
that dropped to 30 in 2008 and 35 in 2009. Over the five years, there were 277 different lotteries – representing 116 unique programs – where admission to a program was determined partially by a random number.

Table 3.2 – Number of Programs with First-Choice Lotteries by Year

<table>
<thead>
<tr>
<th>Year</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lottery</td>
<td>63</td>
<td>73</td>
<td>76</td>
<td>30</td>
<td>35</td>
<td>277</td>
</tr>
<tr>
<td>No Lottery</td>
<td>48</td>
<td>58</td>
<td>72</td>
<td>138</td>
<td>145</td>
<td>461</td>
</tr>
<tr>
<td>Total</td>
<td>111</td>
<td>131</td>
<td>148</td>
<td>168</td>
<td>180</td>
<td>738</td>
</tr>
</tbody>
</table>

Note: This table shows the total number of limited unscreened programs, broken into those that were oversubscribed and held lotteries among their first choice applicants and those that were not. 2005 refers to the admissions process for entry into high school in September 2005.

To confirm randomization in each of these lotteries, I test their balance across a series of pre-high school covariates. In Table 3.3, I show the balance for a sample school: High School A. Table 3.3 examines eleven different pre-intervention characteristics, all measured when a student was in 8th grade. The first three variables related to student academic outcomes, including reading test scores, math test scores, and attendance. The next three variables reflect student demographics, including the percentage of students who are female, the percentage of students who qualify for free or reduced price lunch, and the percentage of students who are black. The final five variables correspond to data collected through the NYC FITNESSGRAM program, including student weight, height, BMI, and the percentage of students who overweight and obese. Because the NYC FITNESSGRAM program did not exist prior to the 2006/07 school year, no 8th grade data on height and weight is available prior to the 2007 entering 9th grade cohort.
Table 3.3 – Covariate Balance for High School A

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. reading test score in 8th gr. (z)</td>
<td>-0.16</td>
<td>-0.23</td>
<td>0.59</td>
<td>-0.28</td>
<td>-0.29</td>
<td>0.91</td>
<td>-0.27</td>
<td>-0.16</td>
<td>0.53</td>
<td>-0.28</td>
<td>0.07</td>
<td>0.15</td>
<td>-0.31</td>
<td>-0.15</td>
<td>0.34</td>
</tr>
<tr>
<td>Avg. math test score in 8th gr. (z)</td>
<td>-0.32</td>
<td>-0.36</td>
<td>0.82</td>
<td>-0.44</td>
<td>-0.42</td>
<td>0.95</td>
<td>-0.55</td>
<td>-0.47</td>
<td>0.58</td>
<td>-0.37</td>
<td>-0.24</td>
<td>0.60</td>
<td>-0.32</td>
<td>-0.37</td>
<td>0.79</td>
</tr>
<tr>
<td>Avg. attendance rate in 8th gr.</td>
<td>86.3</td>
<td>86.4</td>
<td>0.97</td>
<td>85.4</td>
<td>84.0</td>
<td>0.70</td>
<td>82.1</td>
<td>87.3</td>
<td>0.11</td>
<td>87.8</td>
<td>83.5</td>
<td>0.23</td>
<td>88.9</td>
<td>87.9</td>
<td>0.70</td>
</tr>
<tr>
<td>Percent female</td>
<td>84.4</td>
<td>92.9</td>
<td>0.15</td>
<td>81.8</td>
<td>77.0</td>
<td>0.63</td>
<td>89.7</td>
<td>77.6</td>
<td>0.17</td>
<td>76.6</td>
<td>94.1</td>
<td>0.11</td>
<td>71.9</td>
<td>83.3</td>
<td>0.33</td>
</tr>
<tr>
<td>Percent free/reduced price lunch</td>
<td>66.7</td>
<td>67.1</td>
<td>0.96</td>
<td>72.7</td>
<td>67.8</td>
<td>0.66</td>
<td>79.3</td>
<td>70.7</td>
<td>0.40</td>
<td>79.7</td>
<td>76.5</td>
<td>0.78</td>
<td>76.6</td>
<td>66.7</td>
<td>0.40</td>
</tr>
<tr>
<td>Percent black</td>
<td>44.4</td>
<td>48.6</td>
<td>0.67</td>
<td>31.8</td>
<td>49.4</td>
<td>0.14</td>
<td>37.9</td>
<td>41.4</td>
<td>0.76</td>
<td>34.4</td>
<td>35.3</td>
<td>0.94</td>
<td>42.2</td>
<td>33.3</td>
<td>0.50</td>
</tr>
<tr>
<td>Weight in 8th grade (pounds)</td>
<td>. .</td>
<td>. .</td>
<td>. .</td>
<td>154</td>
<td>150</td>
<td>0.85</td>
<td>140</td>
<td>118</td>
<td>0.12</td>
<td>137</td>
<td>137</td>
<td>0.99</td>
<td>137</td>
<td>137</td>
<td>0.99</td>
</tr>
<tr>
<td>Height in 8th grade (inches)</td>
<td>. .</td>
<td>. .</td>
<td>. .</td>
<td>. .</td>
<td>62.9</td>
<td>63.2</td>
<td>0.79</td>
<td>63.6</td>
<td>62.2</td>
<td>0.24</td>
<td>64.1</td>
<td>62.9</td>
<td>0.33</td>
<td>64.1</td>
<td>62.9</td>
</tr>
<tr>
<td>BMI in 8th grade</td>
<td>. .</td>
<td>. .</td>
<td>. .</td>
<td>. .</td>
<td>27.1</td>
<td>26.3</td>
<td>0.79</td>
<td>24.4</td>
<td>21.4</td>
<td>0.23</td>
<td>23.3</td>
<td>24.5</td>
<td>0.51</td>
<td>23.3</td>
<td>24.5</td>
</tr>
<tr>
<td>Percent overweight in 8th grade</td>
<td>. .</td>
<td>. .</td>
<td>. .</td>
<td>. .</td>
<td>62.5</td>
<td>64.7</td>
<td>0.92</td>
<td>51.5</td>
<td>20.0</td>
<td>0.20</td>
<td>48.8</td>
<td>66.7</td>
<td>0.34</td>
<td>48.8</td>
<td>66.7</td>
</tr>
<tr>
<td>Percent obese in 8th grade</td>
<td>. .</td>
<td>. .</td>
<td>. .</td>
<td>. .</td>
<td>50.0</td>
<td>47.1</td>
<td>0.90</td>
<td>30.3</td>
<td>20.0</td>
<td>0.65</td>
<td>14.6</td>
<td>44.4</td>
<td>0.04</td>
<td>14.6</td>
<td>44.4</td>
</tr>
<tr>
<td>Number of students</td>
<td>52</td>
<td>86</td>
<td>25 100</td>
<td>35</td>
<td>67</td>
<td>79 19</td>
<td>75</td>
<td>21</td>
<td>. .</td>
<td>. .</td>
<td>. .</td>
<td>. .</td>
<td>. .</td>
<td>. .</td>
<td>. .</td>
</tr>
</tbody>
</table>

Note: T = Treatment Group, C = Control Group, P = P-value of a t-test of the equality of the treatment and control means. Variables with (z) in the title are normalized z-scores, with a mean of zero and standard deviation of one for the entire population of test-takers within a given year. The "number of students" represents the total number of students in the lottery, though certain students are missing data for certain variables. 2005 refers to the admissions process for entry into high school in September 2005.

In Table 3.3, “T” represents the treatment group outcome. In the top left corner of the table, we see that in 2005, students who chose High School A first and were admitted by lottery (the treatment group) had an average 8th grade reading test score that was 0.16 standard deviations below the mean for New York City. “C” represents the control group. In the second column, we see that students who chose High School A first and were not admitted by lottery (the control group) had an average 8th grade reading test score that was 0.23 standard deviations below the mean for New York City. In the third column, we see that a two-sample t-test for equality of means has a p-value of 0.59, implying that it is not unlikely we would see this type of treatment-control difference by chance. Looking across all five years, only one of the 45 differences is statistically significant at the 0.05 level, a number consistent with (if slightly below) what one would expect by chance.
This analysis can be expanded to all lotteries that took place between 2005 and 2009 in all programs. Figure 3.2 is a bar chart showing the results for average weight in 8th grade. The horizontal axis has one (narrow) bar for each of the 94 lotteries where both the treatment and control group had at least two observations with 8th grade weight recorded. The vertical axis shows the p-value for a t-test of equality of means. Of the 94 t-tests, six have a p-value less than or equal to 0.05, corresponding to 6.4% of the total. This is consistent with what we would expect by chance, and provides evidence that the treatment and control groups were properly randomized across average weight in 8th grade.

---

53 One needs at least two observations to calculate the t-statistic in Equation 1, since both the treatment and control group must have some variation. Of the 277 total lotteries, 159 are missing 8th grade weight for the treatment and/or control groups, since many lotteries took place prior to regular administration of the FITNESSGRAM test. The remaining 24 lotteries where no p-value was calculated have only a single observation in either the treatment or control group.
Table 3.4 – Fraction of Lotteries with Covariate P-Values Less than or Equal to 0.05.

<table>
<thead>
<tr>
<th>Lotteries with P ≤ 0.05</th>
<th>All Lotteries</th>
<th>Fraction with P ≤ 0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. reading test score in 8th grade (z)</td>
<td>8</td>
<td>232</td>
</tr>
<tr>
<td>Avg. math test score in 8th grade (z)</td>
<td>16</td>
<td>234</td>
</tr>
<tr>
<td>Avg. attendance rate in 8th grade</td>
<td>12</td>
<td>235</td>
</tr>
<tr>
<td>Percent female</td>
<td>12</td>
<td>220</td>
</tr>
<tr>
<td>Percent free/reduced price lunch</td>
<td>8</td>
<td>229</td>
</tr>
<tr>
<td>Percent black</td>
<td>17</td>
<td>229</td>
</tr>
<tr>
<td>Weight in 8th grade (pounds)</td>
<td>6</td>
<td>94</td>
</tr>
<tr>
<td>Height in 8th grade (inches)</td>
<td>6</td>
<td>93</td>
</tr>
<tr>
<td>BMI in 8th grade</td>
<td>5</td>
<td>92</td>
</tr>
<tr>
<td>Percent overweight in 8th grade</td>
<td>6</td>
<td>92</td>
</tr>
<tr>
<td>Percent obese in 8th grade</td>
<td>2</td>
<td>85</td>
</tr>
<tr>
<td>Total</td>
<td>98</td>
<td>1,835</td>
</tr>
</tbody>
</table>

Note: P = P-Value of a t-test of the equality of the treatment and control means. Variables with (z) in the title are normalized z-scores, with a mean of zero and standard deviation of one for the entire population of test-takers within a given year. Each program-year combination is counted separately.

Table 3.4 shows the results for all eleven covariates. For each covariate, close to 5% of the t-tests have p-values less than or equal to 0.05, ranging from 2.4% for “percent obese” to 7.4% for “percent black.” Of the 1,835 t-tests run across all lotteries and all nine covariates, 5.3% have p-values less or equal to than 0.05. This balance provides evidence that the lotteries were properly implemented and that the control group is a valid counterfactual for the treatment group.

3.5 Results

In Table 3.5, I show 25 outcome variables for High School A, the same school used as an example in Table 3.3. The first 20 outcome variables are based on the FITNESSGRAM data, including height, weight, BMI, percent overweight, and percent obese. For each category, I record results in the first through fourth year of high school. The last five variables are student
achievement variables, including the average attendance rate and credits earned during the first year of high school, the average English Regents test score during the third year of high school, and the percentage of students who had graduated by the fourth year of high school (with one graduation rate restricted to the more rigorous Regents diploma).

In Table 3.5, we can see that High School A did not begin administering the FITNESSGRAM test until the 2008/09 school year. Students only have first-year height and weight results for the admissions processes in 2008 and 2009. Students who were admitted to 9th grade in 2005 have only a fourth-year result, since the 2008/09 school year corresponds to their

### Table 3.5 – Outcomes for High School A

<table>
<thead>
<tr>
<th></th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Height (Year 1 of HS)</strong></td>
<td>T</td>
<td>C</td>
<td>P</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>64.0</td>
<td>64.0</td>
<td>0.57</td>
<td>64.0</td>
<td>64.0</td>
</tr>
<tr>
<td><strong>Weight (Year 1 of HS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>140.9</td>
<td>139.5</td>
<td>0.90</td>
<td>140.9</td>
<td>139.5</td>
</tr>
<tr>
<td><strong>BMI (Year 1 of HS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>25.5</td>
<td>25.6</td>
<td>0.96</td>
<td>25.5</td>
<td>25.6</td>
</tr>
<tr>
<td><strong>% Overweight (Year 1 of HS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>27.8</td>
<td>31.6</td>
<td>0.81</td>
<td>27.8</td>
<td>31.6</td>
</tr>
<tr>
<td><strong>BMI (Year 2 of HS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>25.0</td>
<td>30.0</td>
<td>0.80</td>
<td>25.0</td>
<td>30.0</td>
</tr>
<tr>
<td><strong>% Overweight (Year 2 of HS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>33.3</td>
<td>35.7</td>
<td>0.88</td>
<td>33.3</td>
<td>35.7</td>
</tr>
<tr>
<td><strong>BMI (Year 3 of HS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>24.8</td>
<td>24.0</td>
<td>0.67</td>
<td>24.8</td>
<td>24.0</td>
</tr>
<tr>
<td><strong>% Overweight (Year 3 of HS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>40.5</td>
<td>38.5</td>
<td>0.90</td>
<td>40.5</td>
<td>38.5</td>
</tr>
<tr>
<td><strong>BMI (Year 4 of HS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>24.2</td>
<td>23.9</td>
<td>0.90</td>
<td>24.2</td>
<td>23.9</td>
</tr>
<tr>
<td><strong>Height (Year 1 of HS)</strong></td>
<td>T</td>
<td>C</td>
<td>P</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>150.0</td>
<td>152.9</td>
<td>0.68</td>
<td>150.0</td>
<td>152.9</td>
</tr>
<tr>
<td><strong>BMI (Year 2 of HS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>24.7</td>
<td>23.4</td>
<td>0.41</td>
<td>24.7</td>
<td>23.4</td>
</tr>
<tr>
<td><strong>BMI (Year 3 of HS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>25.7</td>
<td>25.8</td>
<td>0.96</td>
<td>25.7</td>
<td>25.8</td>
</tr>
<tr>
<td><strong>BMI (Year 4 of HS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>% Overweight (Year 1 of HS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>52.4</td>
<td>50.0</td>
<td>0.92</td>
<td>52.4</td>
<td>50.0</td>
</tr>
<tr>
<td><strong>% Overweight (Year 2 of HS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>48.9</td>
<td>27.3</td>
<td>0.20</td>
<td>48.9</td>
<td>27.3</td>
</tr>
<tr>
<td><strong>% Overweight (Year 3 of HS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>41.2</td>
<td>58.3</td>
<td>0.29</td>
<td>41.2</td>
<td>58.3</td>
</tr>
<tr>
<td><strong>BMI (Year 4 of HS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>% Overweight (Year 4 of HS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>19.0</td>
<td>33.3</td>
<td>0.43</td>
<td>19.0</td>
<td>33.3</td>
</tr>
<tr>
<td><strong>% Obese (Year 1 of HS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>% Obese (Year 2 of HS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>% Obese (Year 3 of HS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>% Obese (Year 4 of HS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Attendance Rate (Year 1 of HS)</strong></td>
<td>73.9</td>
<td>79.5</td>
<td>0.16</td>
<td>73.9</td>
<td>79.5</td>
</tr>
<tr>
<td><strong>Credits Earned (Year 1 of HS)</strong></td>
<td>10.0</td>
<td>8.7</td>
<td>0.18</td>
<td>10.0</td>
<td>8.7</td>
</tr>
<tr>
<td><strong>Eng. Regents Score (Year 3 of HS)</strong></td>
<td>70.5</td>
<td>65.2</td>
<td>0.21</td>
<td>70.5</td>
<td>65.2</td>
</tr>
<tr>
<td><strong>Graduation Rate (Year 4 of HS)</strong></td>
<td>59.0</td>
<td>50.0</td>
<td>0.41</td>
<td>59.0</td>
<td>50.0</td>
</tr>
<tr>
<td><strong>Regents Grad. Rate (Year 4 of HS)</strong></td>
<td>38.5</td>
<td>37.5</td>
<td>0.93</td>
<td>38.5</td>
<td>37.5</td>
</tr>
<tr>
<td><strong>Number of students</strong></td>
<td>52</td>
<td>86</td>
<td></td>
<td>25</td>
<td>100</td>
</tr>
</tbody>
</table>

Note: T = Treatment Group, C = Control Group, P = P-Value of a t-test of the equality of the treatment and control means. Variables with (z) in the title are normalized z-scores, with a mean of zero and standard deviation of one for the entire population of test-takers within a given year.

The "number of students" represents the total number of students in the lottery, though certain students are missing data for certain variables. 2005 refers to the admissions process for entry into high school in September 2005.
Looking across all FITNESSGRAM and student achievement variables, we see only three p-values that are statistically significant at the 0.05 level. The students admitted by lottery to High School A in 2006 had an average English Regents score three years later of 72.8, compared to 63.1 for those students turned away by the lottery, a difference with a p-value of 0.04. The lottery winners in 2009 had a second-year BMI and percentage obese that was sufficiently different from the lottery losers that the p-values were 0.05 and 0.03, respectively. Aside from these three results, no other p-values on the table are equal to or below 0.05. Seeing three of 58 p-values equal to or below 0.05 is roughly what one would expect by chance if High School A had no differential effect on FITNESSGRAM or student achievement outcomes.

As with the tests for covariate balance in the prior section, we can repeat the analysis for all programs that had at least one lottery between 2005 and 2009. The results for one variable – average weight three years after a student starts high school – are shown in Figure 3.3. As before, the height of each bar represents the p-value for a t-test of equality of means for one of the 123 lotteries held between 2005 and 2009. Six of these lotteries, representing 4.9% of the total, had p-values equal to or lower than 0.05. These results are consistent with what one would expect to see if the null hypothesis were true, and schools have no differential effect on student weight in their third year of high school.
Table 3.6 shows the results for all lotteries across each of the 25 different outcome variables. Overall, we see little evidence of differential school effects on variables related to student weight. Looking at BMI, 4.1% of lotteries have p-values at or below 0.05, close to what one would expect by chance. The results for weight, percentage overweight, and percentage obese are similar. Looking at student height, where I would a priori expect schools to have little impact, we see this to be the case. Out of 370 lotteries, 6.8% have p-values less than or equal to 0.05.
Table 3.6 – Fraction of Lotteries with Outcome P-Values Less than or Equal to 0.05

<table>
<thead>
<tr>
<th></th>
<th>Lotteries with</th>
<th>Fraction with P ≤ 0.05</th>
<th>All Lotteries</th>
<th>≤ 0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Height</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Height (Year 1 of HS)</td>
<td>5</td>
<td>62</td>
<td>0.081</td>
<td></td>
</tr>
<tr>
<td>Height (Year 2 of HS)</td>
<td>5</td>
<td>104</td>
<td>0.048</td>
<td></td>
</tr>
<tr>
<td>Height (Year 3 of HS)</td>
<td>7</td>
<td>123</td>
<td>0.057</td>
<td></td>
</tr>
<tr>
<td>Height (Year 4 of HS)</td>
<td>8</td>
<td>81</td>
<td>0.099</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>370</td>
<td>0.068</td>
<td></td>
</tr>
<tr>
<td><strong>Weight</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight (Year 1 of HS)</td>
<td>0</td>
<td>61</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Weight (Year 2 of HS)</td>
<td>3</td>
<td>104</td>
<td>0.029</td>
<td></td>
</tr>
<tr>
<td>Weight (Year 3 of HS)</td>
<td>6</td>
<td>123</td>
<td>0.049</td>
<td></td>
</tr>
<tr>
<td>Weight (Year 4 of HS)</td>
<td>1</td>
<td>80</td>
<td>0.013</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>368</td>
<td>0.027</td>
<td></td>
</tr>
<tr>
<td><strong>Body Mass Index</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMI (Year 1 of HS)</td>
<td>1</td>
<td>61</td>
<td>0.016</td>
<td></td>
</tr>
<tr>
<td>BMI (Year 2 of HS)</td>
<td>6</td>
<td>103</td>
<td>0.058</td>
<td></td>
</tr>
<tr>
<td>BMI (Year 3 of HS)</td>
<td>6</td>
<td>123</td>
<td>0.049</td>
<td></td>
</tr>
<tr>
<td>BMI (Year 4 of HS)</td>
<td>2</td>
<td>79</td>
<td>0.025</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>366</td>
<td>0.041</td>
<td></td>
</tr>
<tr>
<td><strong>% Overweight</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Overweight (Year 1 of HS)</td>
<td>2</td>
<td>59</td>
<td>0.034</td>
<td></td>
</tr>
<tr>
<td>% Overweight (Year 2 of HS)</td>
<td>6</td>
<td>102</td>
<td>0.059</td>
<td></td>
</tr>
<tr>
<td>% Overweight (Year 3 of HS)</td>
<td>7</td>
<td>119</td>
<td>0.059</td>
<td></td>
</tr>
<tr>
<td>% Overweight (Year 4 of HS)</td>
<td>2</td>
<td>77</td>
<td>0.026</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>17</td>
<td>357</td>
<td>0.048</td>
<td></td>
</tr>
<tr>
<td><strong>% Obese</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Obese (Year 1 of HS)</td>
<td>2</td>
<td>53</td>
<td>0.038</td>
<td></td>
</tr>
<tr>
<td>% Obese (Year 2 of HS)</td>
<td>5</td>
<td>98</td>
<td>0.051</td>
<td></td>
</tr>
<tr>
<td>% Obese (Year 3 of HS)</td>
<td>4</td>
<td>110</td>
<td>0.036</td>
<td></td>
</tr>
<tr>
<td>% Obese (Year 4 of HS)</td>
<td>0</td>
<td>68</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>329</td>
<td>0.033</td>
<td></td>
</tr>
<tr>
<td><strong>Student Achievement Measures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attendance Rate (Year 1 of HS)</td>
<td>18</td>
<td>227</td>
<td>0.079</td>
<td></td>
</tr>
<tr>
<td>Credits Earned (Year 1 of HS)</td>
<td>34</td>
<td>226</td>
<td>0.150</td>
<td></td>
</tr>
<tr>
<td>English Regents Score (Year 3 of HS)</td>
<td>17</td>
<td>180</td>
<td>0.094</td>
<td></td>
</tr>
<tr>
<td>Graduation Rate (Year 4 of HS)</td>
<td>9</td>
<td>111</td>
<td>0.081</td>
<td></td>
</tr>
<tr>
<td>Regents Graduation Rate (Year 4 of HS)</td>
<td>8</td>
<td>107</td>
<td>0.075</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>86</td>
<td>851</td>
<td>0.101</td>
<td></td>
</tr>
</tbody>
</table>

Note: P = P-Value of a t-test of the equality of the treatment and control means. Each program-year combination is counted separately.
The results are somewhat different for academic outcomes. This can be seen most clearly in Figure 3.4, which is a graphical representation of the last column of Table 3.6. The figure is sorted in ascending order, based on the percentage of p-values that are equal to or less than 0.05. The darker bars correspond to the FITNESSGRAM measures, which generally had about 5% of p-values at or below the 0.05 level. Those that are the farthest from 5% -- either above or below -- correspond to FITNESSGRAM measures from Year 1 or Year 4 of high school -- where the sample size of lotteries is lowest (see Table 3.6).

Figure 3.4 - Fraction of Lotteries with Outcome P-Values Less than or equal to 0.05

The lighter bars correspond to student achievement outcomes. For all five outcomes shown here, the treatment group and control group were statistically different from one another at the 0.05 level more than 5% of the time. The largest effect is for credits earned in the first
year of high school, where 34 of the 226 lotteries had p-values at or below 0.05, corresponding to 15% of the total. Given that schools might be expected to influence student credits through two channels – both the quality of instruction and the ease with which they award credits – it is perhaps not surprising that differential school effects show up most frequently on this measure.

3.6 Conclusion and Discussion

In this study, I assess the impact of attending one New York City high school versus another on student weight and weight-related outcomes. For a group of New York City high schools that admit students by lottery, I compare the weight-related outcomes for two groups of students who chose the school first. The first group – the treatment group – are those who chose the school first and were admitted based purely on a random number. The second group – the control group – are those who chose the school first and were not admitted based purely on a random number. Overall, by tabulating the results for hundreds of small lotteries, I find little evidence that schools have a differential impact on student weight or weight-related variables.

There are several limitations to this study. First, I can make no claim about the overall effect of going to high school in New York City on student weight loss or gain. There may be system-wide programs in New York City that have a substantial effect on student obesity rates, and this study is not designed to capture their effectiveness. Rather, I look specifically at whether there is a differential effect of being admitted to a particular school as one’s first choice, rather than being admitted to one of the various schools chosen lower on the application (or, in some cases, a school that a student did not choose).

Second, my counterfactual for each school is somewhat different, reflecting the preferences of the students who apply to it. To the extent that students list similar schools as their
first and lower choices, the schools attended by the control group may be similar to the schools attended by the treatment group. Therefore, it is possible that differential school effects exist, but only when comparing schools that are less alike than those a student lists on her application.

Third, in contrast to some research cited above, all of the schools in this study are part of a single school district. It may be that some of the school effects found in prior research are actually district effects. The quality of a school’s lunches or a school’s propensity to install snack and soda machines may vary more by district than by school, limiting my ability to capture their impact in this study. At the same time, individual schools in New York do undertake obesity reduction programs, and other factors such as the number of sports teams offered or the body composition of the incoming class – do vary by school.

Fourth, most of my lotteries are small, sometimes very small. High School A, shown as an example in Tables 3.2 and 3.4, has at least 19 students in both the treatment and control group for all five years of the study. However, it is the exception, rather than the rule. Of the 277 lotteries shown in Table 3.2, only 41 had 19 or more students in both the treatment and control group (five of which were for High School A). Any individual lottery, therefore, would be unlikely to have a p-value less than 0.05 unless the school effect was particularly large. By combining lotteries as I do in Table 3.6, I begin to address this concern. Even with a small school effect, one would expect to see more than 5% of lotteries have a p-value below 0.05. However, with small lotteries, it might not be much greater than 5%.

---

54 This study combines lotteries by reviewing each one individually and assessing the proportion that have p-values less than 0.05. Another possibility would be to pool all lotteries into a single large lottery. I believe this is not appropriate for the current question, because I have no reason to expect that the treatment schools have systematically higher (or lower) weight outcomes than the control schools. Rather, I am testing whether they have systematically different weight outcomes. If certain treatment schools have lower weight outcomes than their control schools (perhaps because of an effective exercise program), and certain treatment schools have higher weight outcomes than their control schools (perhaps because they are located next to a fast food restaurant), these effects might cancel out in a pooled sample. However, using my methodology, I should capture both of these effects, as each is likely to generate p-values below 0.05 more often than one would expect by chance.
Given the small size of the lotteries – which make detecting any effect difficult – the student achievement results are all the more remarkable. Fifteen percent of the lotteries show a statistically significant effect at the 0.05 level on 9th grade credits earned, despite the fact that the vast majority of lotteries have fewer than 20 students in either the treatment or control group. The other achievement results show a similar – though less dramatic – pattern, with more than 5% of p-values below 0.05. In future work, I plan to expand on these findings, looking at whether the academic results are consistent in one direction over time – i.e. schools have consistently positive or negative effects – and assessing how they align with common non-experimental methods of evaluating schools. For the present study, however, the student achievement results make a clear point: even with small lotteries, differential school effects are detectable.
4. Appendix

Table A.1.1 – Subjective Division of Teacher Survey Questions

**Intrinsic Motivation Questions (Subjective Division)**

<table>
<thead>
<tr>
<th>Number</th>
<th>Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>t_q2b</td>
<td>Teachers in this school set high standards for student work in their classes</td>
</tr>
<tr>
<td>t_q5a</td>
<td>To what extent do you feel supported by: your principal</td>
</tr>
<tr>
<td>t_q5c</td>
<td>To what extent do you feel supported by: other teachers at your school</td>
</tr>
<tr>
<td>t_q6b</td>
<td>School leaders invite teachers to play a meaningful role in setting goals and making important decisions for this school</td>
</tr>
<tr>
<td>t_q6d</td>
<td>Teachers in my school respect teachers who take the lead in school improvement efforts</td>
</tr>
<tr>
<td>t_q6e</td>
<td>Teachers in my school trust each other</td>
</tr>
<tr>
<td>t_q6f</td>
<td>Teachers in my school recognize and respect colleagues who are the most effective teachers</td>
</tr>
<tr>
<td>t_q6h</td>
<td>School leaders give me regular and helpful feedback about my teaching</td>
</tr>
<tr>
<td>t_q6i</td>
<td>School leaders place a high priority on the quality of teaching at this school</td>
</tr>
</tbody>
</table>

**Other Questions (Subjective Division)**

<table>
<thead>
<tr>
<th>Number</th>
<th>Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>t_q1a</td>
<td>School leaders communicate a clear vision for this school</td>
</tr>
<tr>
<td>t_q1c</td>
<td>School leaders encourage open communication on important school issues</td>
</tr>
<tr>
<td>t_q1d</td>
<td>Curriculum, instruction, and assessment are aligned within and across the grade levels at this school</td>
</tr>
<tr>
<td>t_q1e</td>
<td>The principal places the learning needs of children ahead of other interests</td>
</tr>
<tr>
<td>t_q1f</td>
<td>The principal is an effective manager who makes the school run smoothly</td>
</tr>
<tr>
<td>t_q1g</td>
<td>I trust the principal at his or her word</td>
</tr>
<tr>
<td>t_q2c</td>
<td>My school has clear measures of progress for student achievement throughout the year</td>
</tr>
<tr>
<td>t_q2d</td>
<td>This school makes it a priority to help students develop challenging learning goals</td>
</tr>
<tr>
<td>t_q2e</td>
<td>This school makes it a priority to help students find the best ways to achieve their learning goals</td>
</tr>
<tr>
<td>t_q4</td>
<td>My school offers a wide enough variety of courses to keep students at my school engaged</td>
</tr>
<tr>
<td>t_q6a</td>
<td>The principal has confidence in the expertise of the teachers</td>
</tr>
<tr>
<td>t_q6c</td>
<td>School leaders provide time for collaboration among teachers</td>
</tr>
<tr>
<td>t_q6g</td>
<td>School leaders visit classrooms to observe the quality of teaching at this school</td>
</tr>
<tr>
<td>Question</td>
<td>Description</td>
</tr>
<tr>
<td>----------</td>
<td>-------------</td>
</tr>
<tr>
<td>t_q6k</td>
<td>Most teachers in my school work together on teams to improve their instructional practice</td>
</tr>
<tr>
<td>t_q6l</td>
<td>Teachers in my school use student achievement data to improve instructional decisions</td>
</tr>
<tr>
<td>t_q7a</td>
<td>The professional development I received this year provided me with teaching strategies to better meet the needs of my students</td>
</tr>
<tr>
<td>t_q7b</td>
<td>I have sufficient materials to teach my class(es), including: books, audio/visual equipment, maps, and/or calculators</td>
</tr>
<tr>
<td>t_q7d</td>
<td>This year, I received helpful training on the use of student achievement data to improve teaching and learning</td>
</tr>
<tr>
<td>t_q7e</td>
<td>The professional development I received this year provided me with content support in my subject area</td>
</tr>
<tr>
<td>t_q8a</td>
<td>Obtaining information from parents about student learning needs is a priority at my school</td>
</tr>
<tr>
<td>t_q8b</td>
<td>Teachers and administrators in my school use information from parents to improve instructional practices and meet student learning needs</td>
</tr>
<tr>
<td>t_q8c</td>
<td>My school communicates effectively with parents when students misbehave</td>
</tr>
<tr>
<td>t_q10d</td>
<td>How often during this school year have you: sent parents written information on what you are teaching and what students are expected to learn</td>
</tr>
<tr>
<td>t_q10e</td>
<td>How often during this school year have you: sent home information on services to help students or parents such as: tutoring, after-school programs, or workshops adults can attend to help their children in school</td>
</tr>
<tr>
<td>t_q11a</td>
<td>Order and discipline are maintained at my school</td>
</tr>
<tr>
<td>t_q11b</td>
<td>I can get the help I need at my school to address student behavior and discipline problems</td>
</tr>
<tr>
<td>t_q11c</td>
<td>I am safe at my school</td>
</tr>
<tr>
<td>t_q11d</td>
<td>Crime and violence are a problem in my school</td>
</tr>
<tr>
<td>t_q11e</td>
<td>Students in my school are often threatened or bullied</td>
</tr>
<tr>
<td>t_q11g</td>
<td>Most students at my school treat teachers with respect</td>
</tr>
<tr>
<td>t_q11h</td>
<td>Most parents treat teachers at this school with respect</td>
</tr>
<tr>
<td>t_q11j</td>
<td>Students' use of alcohol and illegal drugs in school is a problem at my school</td>
</tr>
<tr>
<td>t_q11k</td>
<td>There are conflicts at my school based on race, color, creed, ethnicity, national origin, citizenship/immigration status, religion, gender, gender identity, gender expression, sexual orientation or disability</td>
</tr>
<tr>
<td>t_q11l</td>
<td>There is a person or a program in my school that helps students resolve conflicts</td>
</tr>
<tr>
<td>t_q11m</td>
<td>Gang activity is a problem in my school</td>
</tr>
<tr>
<td>t_q11n</td>
<td>My school is kept clean</td>
</tr>
</tbody>
</table>
Table A.1.2 - Intrinsic Motivation Questions vs. Other Questions (Subjective Division, Controls for 2007)

<table>
<thead>
<tr>
<th></th>
<th>Pre-Pgm.</th>
<th>Program</th>
<th>Post-Program</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>( D_i ) Coefficient</td>
<td>2007</td>
<td>2008</td>
<td>2009</td>
</tr>
<tr>
<td>Intrinsic Motivation Questions</td>
<td>N/A</td>
<td>0.07</td>
<td>0.17*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.08)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>Other Questions</td>
<td>N/A</td>
<td>0.08</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.08)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>Difference (Intrinsic - Other)</td>
<td>N/A</td>
<td>-0.01</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.04)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>N</td>
<td>N/A</td>
<td>299</td>
<td>299</td>
</tr>
</tbody>
</table>

Note: each coefficient is the result of a separate regression of the mean school-level teacher survey scores as the dependent variable and treatment \( (D_i) \) as the independent variable. Regressions are weighted by the number of responders at each school. Robust standard errors in parentheses. *** \( p<0.01 \), ** \( p<0.05 \), * \( p<0.1 \)
Table A.1.3 - Intrinsic Motivation Questions vs. Other Questions (Factor Analysis Division, Controls for 2007)

<table>
<thead>
<tr>
<th></th>
<th>Pre-Pgm.</th>
<th>Program</th>
<th>Post-Program</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>( D_i ) Coefficient</td>
<td>2007</td>
<td>2008</td>
<td>2009</td>
</tr>
<tr>
<td>Intrinsic Motivation Questions</td>
<td>N/A</td>
<td>0.11</td>
<td>0.21*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.10)</td>
<td>(0.12)</td>
</tr>
<tr>
<td>School Safety Questions</td>
<td>N/A</td>
<td>0.01</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.08)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>Difference (Intrinsic - Safety)</td>
<td>N/A</td>
<td>0.14</td>
<td>0.17*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.09)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>N</td>
<td>N/A</td>
<td>299</td>
<td>299</td>
</tr>
</tbody>
</table>

Note: each coefficient is the result of a separate regression of the mean school-level teacher survey scores as the dependent variable and treatment \( (D_i) \) as the independent variable. Regressions are weighted by the number of responders at each school. Robust standard errors in parentheses. *** \( p<0.01 \), ** \( p<0.05 \), * \( p<0.1 \)
Table A.1.4 – Teacher Survey Triple Difference Regression

<table>
<thead>
<tr>
<th></th>
<th>Pre-Pgm.</th>
<th>Program</th>
<th>Post-Program</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>$D_i$ Coefficient</td>
<td>2007</td>
<td>2008</td>
<td>2009</td>
</tr>
<tr>
<td>Subjective Division</td>
<td>N/A</td>
<td>-0.01</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.04)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>Factor Analysis Division</td>
<td>N/A</td>
<td>0.16</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.11)</td>
<td>(0.12)</td>
</tr>
<tr>
<td>N</td>
<td>N/A</td>
<td>302</td>
<td>302</td>
</tr>
</tbody>
</table>

Note: each coefficient is the result of a separate regression. The dependent variable is a difference-in-difference variable based on school-level survey scores. The first difference is between survey scores related to intrinsic motivation and those unrelated. The second difference is between the year shown and 2007. The independent variable is a dummy ($D_i$) indicating whether or not the school was randomly offered treatment. Regressions are weighted by the number of responders at each school. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1
Table A.1.5 – Impact of Bonus Eligibility on Student Test Scores (RD, Local Linear Regression)

<table>
<thead>
<tr>
<th></th>
<th>Pre-Program</th>
<th>Program</th>
<th>Post-Program</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>$E_s$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coefficient</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2006</td>
<td>2007</td>
<td>2008</td>
</tr>
<tr>
<td>Math, Treatment</td>
<td>0.05</td>
<td>0.08</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>(0.08)</td>
<td>(0.08)</td>
<td>(0.08)</td>
</tr>
<tr>
<td>ELA, Treatment</td>
<td>0.05</td>
<td>0.02</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.07)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>Math, Control</td>
<td>0.13*</td>
<td>0.14</td>
<td>0.15*</td>
</tr>
<tr>
<td></td>
<td>(0.08)</td>
<td>(0.09)</td>
<td>(0.08)</td>
</tr>
<tr>
<td>ELA, Control</td>
<td>0.13*</td>
<td>0.08</td>
<td>0.15**</td>
</tr>
<tr>
<td></td>
<td>(0.08)</td>
<td>(0.07)</td>
<td>(0.07)</td>
</tr>
</tbody>
</table>

Note: Each number is the coefficient on $E_s$ in a separate local linear regression of school-level mean test score on distance from the eligibility cut point and an indicator for bonus eligibility ($E_s$). The regressions are fit separately on either side of the cut point, and are weighted by the number of test takers in each school. Robust standard errors are in parentheses. The sample size is the same in each year because only schools with testing data in each year were included. *** p<0.01, ** p<0.05, * p<0.1
Table A.2.1 – NHIS/Medicare Data Linkage

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NHIS 1994-1998</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NHIS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicare</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NHIS 1999-2005</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NHIS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicare</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1999</td>
<td>2000</td>
<td>2001</td>
<td>2002</td>
<td>2003</td>
<td>2004</td>
<td>2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: the Medicare years labeled on the chart are potentially useful for our study because they represent a Medicare record that is linked to a later NHIS interview. Medicare records linked to earlier NHIS interviews provide no information on quitting behavior after CAD treatment.

For those respondents who were diagnosed with CAD prior to their NHIS interview date, we have the ability to look at their smoking behavior before and after their diagnosis. For the subset of CAD patients who underwent PCI or CABG, we can also look at their smoking behavior before and after their procedure. For example, for individuals interviewed in 1994 who had PCI, we can look at their smoking behavior before and after their procedure only if they underwent PCI between 1991 and 1994 (and within 1994, only if their procedure was before the date of the NHIS interview). If a person had PCI before 1991, then we have no record of their procedure. If a person had PCI after 1994, then we have no record of their smoking behavior after their procedure.

Each person in the linked dataset, therefore, has a “diagnosis window” within which they must be diagnosed with CAD to be included in our study. The longest window is for a person who was interviewed in 1998 – she will be included if she was enrolled in Medicare and diagnosed with CAD between 1991 and 1998 (prior to the interview date). The shortest window is for a person who was interviewed in early 1999 – she will be included only if she was enrolled in Medicare and diagnosed with CAD on an earlier date in 1999 than the date of her interview.
Table A.2.2 – Regression Results for Discrete Time Linear Probability Hazard Model, Testing for Significance of Pre-Treatment Differences in Quit Rate

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.048***</td>
<td>0.048***</td>
<td>0.047***</td>
<td>0.045***</td>
<td>-0.048</td>
</tr>
<tr>
<td>PCI * Period 7</td>
<td>0.003</td>
<td>0.003</td>
<td>0.005</td>
<td>0.005</td>
<td>0.013</td>
</tr>
<tr>
<td>PCI * Period 8</td>
<td>-0.012</td>
<td>-0.012</td>
<td>-0.013</td>
<td>-0.013</td>
<td>-0.014</td>
</tr>
<tr>
<td>PCI * Period 9</td>
<td>0.037</td>
<td>0.035</td>
<td>0.030</td>
<td>0.029</td>
<td>0.031</td>
</tr>
<tr>
<td>PCI * ∆After</td>
<td>0.091**</td>
<td>0.055</td>
<td>0.036</td>
<td>0.035</td>
<td>0.023</td>
</tr>
<tr>
<td>CABG * Period 7</td>
<td>0.027</td>
<td>0.027</td>
<td>0.027</td>
<td>0.024</td>
<td>0.029</td>
</tr>
<tr>
<td>CABG * Period 8</td>
<td>0.023</td>
<td>0.022</td>
<td>0.018</td>
<td>0.016</td>
<td>0.011</td>
</tr>
<tr>
<td>CABG * Period 9</td>
<td>0.057*</td>
<td>0.054*</td>
<td>0.048</td>
<td>0.045</td>
<td>0.045</td>
</tr>
<tr>
<td>CABG * ∆After</td>
<td>0.226***</td>
<td>0.197***</td>
<td>0.161***</td>
<td>0.156***</td>
<td>0.137***</td>
</tr>
<tr>
<td>∆Diagnosed</td>
<td>0.049***</td>
<td>0.041***</td>
<td>0.025***</td>
<td>0.025***</td>
<td>-0.007</td>
</tr>
<tr>
<td>∆AMI</td>
<td>0.081***</td>
<td>0.074***</td>
<td>0.074***</td>
<td>0.075***</td>
<td></td>
</tr>
</tbody>
</table>

| CABG * Period 7 - PCI * Period 7 | 0.024 | 0.024 | 0.022 | 0.019 | 0.016 |
| CABG * Period 8 - PCI * Period 8 | 0.035 | 0.034 | 0.031 | 0.029 | 0.025 |
| CABG * Period 9 - PCI * Period 9 | 0.020 | 0.019 | 0.018 | 0.016 | 0.014 |
| CABG * ∆After - PCI * ∆After | 0.135** | 0.142** | 0.125** | 0.121** | 0.114** |

| Elixhauser | X | X | X |
| Demographics | X | X | |
| Year, Period, Age, Yrs Smoked | X | |
| Observations | 26,658 | 26,658 | 26,658 | 26,658 | 26,658 |
| Individuals | 3,065 | 3,065 | 3,065 | 3,065 | 3,065 |

Robust standard errors, clustered at the individual level, in parentheses. Regressions are weighted by NHIS probability weights. ∆AMI is an indicator for a patient having her first AMI (in our data) in a particular period. Specifications 3 and 4 include 29 dummy variables indicating when the Elixhauser comorbidity conditions were first diagnosed (in our data). In specification 4, demographic controls include gender, race, education dummies, and income category dummies (including a dummy for missing income data). In specification 5, we also include year dummies, period dummies, age and age squared, and a variable indicating the number of years a person had been smoking as of the baseline period (9.5 years before treatment). *CABG * ∆After - PCI * ∆After* is the difference between the PCI * ∆After and CABG * ∆After coefficients. * significant at the 0.10 level; ** significant at the 0.05 level; *** significant at the 0.01 level.
5. Bibliography


Chen, Tao, Wei Li, Yang Wang, Bo Xu, and Jin Guo. 2012. “Smoking Status on Outcomes after Percutaneous Coronary Intervention.” Clinical Cardiology 35 (9): 570-574.


Deci, E. and Ryan, R (2014): Intrinsic Motivation Inventory. Downloaded from http://www.selfdeterminationtheory.org/ on 7/14/2014


