2017

Essays In Public Economics

Michael Chirico
University of Pennsylvania, chiricom@sas.upenn.edu

Essays In Public Economics

Abstract
This dissertation consists of three chapters on topics in public economics. The first chapter examines the labor market for public school teachers in Wisconsin. By stitching together publicly available cross-sectional data to form a 20-year panel of teachers, I am able to replicate and extend the work of Hanushek, Kain and Rivkin who performed a similar analysis in Texas. The main takeaway is that teachers appear to select on wages, but that student characteristics appear more important in predicting teacher churn. In the second chapter, I present short-term analysis of a randomized-controlled trial designed to test the efficacy of active learning methods for teaching intermediate calculus to first-year college students. The results were inconclusive, suggesting substantial heterogeneity in student preferences and aptitudes for different styles of learning. The final chapter presents the analysis of a large-scale randomized-controlled trial evaluating the potential for messaging-based nudges to elicit increased real estate tax compliance in Philadelphia. Our primary conclusions are that most proposed messaging strategies are indistinguishable from a plainly-worded reminder bill (the exception being consequentialist letters threatening repercussive action absent compliance), but that the saliency per se of a plainly-worded bill can induce late payers to remunerate more quickly.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Economics

First Advisor
Holger Sieg

Keywords
taxation

Subject Categories
Economics | Education

ESSAYS IN PUBLIC ECONOMICS

Michael Chirico

A DISSERTATION
in
Economics

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

Holger Sieg, J. M. Cohen Term
Chair in Economics

Graduate Group Chairperson

Jesus Fernandez-Villaverde, Professor of Economics

Dissertation Committee:
Holger Sieg, J. M. Cohen Term Chair in Economics
Hanming Fang, Class of 1965 Term Professor of Economics
Petra Todd, Edmund J. and Louise W. Kahn Term Professor of Economics
Matthew Steinberg, Assistant Professor of Education

ESSAYS IN PUBLIC ECONOMICS

To you

Acknowledgments

It is with deep reverence for their patience, persistence and perspicacity that I extend my gratitude first and foremost to my committee: Holger Sieg, Hanming Fang, Matthew Steinberg, and Petra Todd. I have also benefitted enormously professionally from the guidance and insight of other coauthors: Rebecca Maynard, Charles Loeffler, Robert Inman, John MacDonald, and Seth Flaxman.

My progression as a researcher also owes an eternal debt to my ever-inquisitive workplace proximity associates read: dearest friends, Pau Pereira Batlle, Gustavo Camilo, Wendy Castillo, Alberto Ciancio, and Juan-Manuel Hernandez. My continued (relative) mental stability is due to them as well as to my other wonderful friends I have made in this city I love – Daniel Wills, Constanza Vergara, Ana Gazmuri, Alix Barasch, Cinthia Konichi, Pedro Olea, and Duna Gylfadottir.

Before the music starts I also need to give a big shoutout of profound unconditional love to my ever-supportive beautiful mess of a family – Ma, Mike, Dad, Sharon, Grandma, Grandpa, Chris, Alyssa, Mark, Brett, Jeremy, Dev, Bryana, y'all're crazy and I love every minute of it.

The research presented here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305B090015 to the University of Pennsylvania. The opinions expressed are those of the author and do not represent the views of the Institute or the U.S. Department of Education.

Michael Chirico

Philadelphia, PA

July 19, 2017

ABSTRACT

ESSAYS IN PUBLIC ECONOMICS

Michael Chirico

Holger Sieg

This dissertation consists of three chapters on topics in public economics. The first chapter examines the labor market for public school teachers in Wisconsin. By stitching together publicly available cross-sectional data to form a 20-year panel of teachers, I am able to replicate and extend the work of Hanushek, Kain and Rivkin who performed a similar analysis in Texas. The main takeaway is that teachers appear to select on wages, but that student characteristics appear more important in predicting teacher churn. In the second chapter, I present short-term analysis of a randomized-controlled trial designed to test the efficacy of active learning methods for teaching intermediate calculus to first-year college students. The results were inconclusive, suggesting substantial heterogeneity in student preferences and aptitudes for different styles of learning. The final chapter presents the analysis of a large-scale randomized-controlled trial evaluating the potential for messaging-based nudges to elicit increased real estate tax compliance in Philadelphia. Our primary conclusions are that most proposed messaging strategies are indistinguishable from a plainly-worded reminder bill (the exception being consequentialist letters threatening repercussive action absent compliance), but that the saliency *per se* of a plainly-worded

bill can induce late payers to remunerate more quickly.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This work is dedicated to understanding and applying a wide range of topics in public economics. In three chapters, I explore public elementary and high school teachers' revealed preferences for pecuniary and non-pecuniary aspects of compensation; the potential for novel technology-assisted pedagogical modes to strengthen the STEM pipeline at American undergraduate institutions; and the efficacy of low-cost, behaviorally-founded tools for local governments to recoup their debts quickly.

The first chapter is all about investigating the labor market for teachers and which incentives can be targeted by policymakers to improve student outcomes. Good teachers have large impacts on student achievement[1]. It is therefore imperative for public schools to be able to attract and retain high-quality teachers. Of preeminent concern for policymakers, then, is the strength of the various manipulable levers at their

[1]See, e.g., Rockoff ([2004]).

disposal for influencing teacher labor markets. More specifically, state and local education administrators would like to identify the policy implications of various tools on three types of teacher mobility: intra-district switching, where due to the collectively bargained nature of most teachers' salaries, only nonpecuniary considerations matter, inter-district switching, where teachers move to another school district in the same state, and exo-district switching, where teachers leave the public teaching workforce entirely[2].

The brunt of this chapter is a replication in a new context (Wisconsin) and time horizon (2000 - 2010) of Hanushek, Kain, and Rivkin ([2004]), who analyze various predictors of teacher churn. The headline results of Hanushek, Kain, and Rivkin ([2004]) were that "teacher mobility is much more strongly related to characteristics of the students, particularly race and achievement, than to salary, although salary exerts a modest impact once compensating differentials are taken into account." I confirm the pith of this conclusion, namely that student characteristics are a much better predictor of turnover than are wage differentials, though I come to different conclusions regarding more specific points. To wit, while I do find strong evidence that the socioeconomic makeup of a teacher's district predicts turnover (and that there is heterogeneity in this effect by teacher race), the evidence I find for the importance of

[2]Policies that affect the supply and quality of new teachers to the profession may also be of considerable importance to replenishing and improving the stock of teachers over time, but I do not consider these channels in this work. See Harris and Sass ([2011]), Wayne et al. ([2008]), and Boyd et al. ([2009]).

wages and student achievement is far from compelling.

I explore to the extent possible potential contributors to this discrepancy in results; most salient are the measurable differences between Texas, where Hanushek, Kain, and Rivkin ([2004]) conduct their study, and Wisconsin. Wisconsin is a largely rural state – its largest city/metropolitan area, Milwaukee, currently has roughly 600,000 residents (1,500,000 including the metropolitan area), making it around the 30th-largest city in the United States[3]. By contrast, Texas has six cities larger than this, with El Paso (#6 in Texas) being the nearest in size to Milwaukee. Though the non-urban parts of Texas are themselves sparsely populated and distinctly rural, the more uniform lack of major population centers in Wisconsin is likely to be reflected when considerably different preferences among local vis--vis urban residents for various aspects of potential teaching positions are aggregated.

To the end of exploring the pecuniary aspects of teacher turnover, I start by expanding upon the efforts of Hanushek, Kain, and Rivkin ([2004]) to infer teachers' tenure-wage paths from teacher-level data on pay. Most unionized teachers are paid according to a salary schedule (specifying wages as an increasing function of tenure and certification) explicated in contracts collectively bargained at the district level. With this easily-obtained information in hand, teachers are able to infer their future potential wage trajectories at their own and other potential district employers. Lacking the physical contract faced by the teachers, an econometrician armed only with

[3]In fact, Milwaukee is the only city in Wisconsin considered to be "large" for NCES reporting purposes.

administrative data reporting actual wages in a given year must use some imputation techniques to deduce the underlying wage structure. I explore the utility of natural Constrained B-Splines (COBS) to this end. COBS are an enhanced version of the traditional semiparametric splines technique enhanced by the ability to impose a monotonicity constraint on the resultant curve which allows the fit to incorporate more local information from nearby experience cells.

The fidelity and utility of the resulting fitted contract curves are supreme. In both large and small districts, COBS produces a plausible tenure-wage arc which enables us to examine counterfactual wage levels for mobile teachers. By comparing the fit to a small number of wage tables obtained from actual contracts, I also learn that using COBS may be preferable to an attempt to use actual wage tables, as the data-derived curves can reveal latent progress of teachers towards further certification, an aspect which is commonly observed in salary tables but rarely included in teacher-level data.

The second Chapter continues to have improving student-teacher interactions as a motivation, but proceeds to study an older segment of the education sector – college. With the rise of technology in the classroom have come a variety of approaches to teaching course material based on methods unavailable in the past due to absent technological tools. While the literature formally aiming to identify causal effects of these new methods on student outcomes is growing, there is still as yet no broad conclusions about if, when, and how such new methods can be used to serve the needs of students and/or simplify the process of learning for students and instructors alike.

This chapter seeks to advance our understanding of factors that influence students' engagement and learning in college mathematics. The particular focus of the study is a comparison of traditional and active instructional methods (with online components) in the context of an intermediate calculus course at a mid-sized private university in the Northeast. In theory, the appeal of active learning classrooms is that students taught in this mode will more fully engage with the curriculum and, as a result, that they will attain deeper understanding and mastery of mathematical concepts, thereafter proceeding to pursue degrees in math or the sciences

I test this hypothesis by block-randomizing students between the two pedagogical modes and monitoring their performance in the course. I supplement the quantitative measures of student progress with qualitative data obtained through formal course observations designed to elicit an understanding of the differences in learning environments and treatment fidelity among the six lecture sections.

Given substantial noncompliance observed in the data, I use the potential outcomes framework of, e.g., Rubin ([1974]), G. W. Imbens and Rubin ([2015]) and Angrist and Imbens ([1995]) to construct intent-to-treat (ITT) and local average treatment effect (LATE) estimates of active learning environments on student performance. I incorporate the blocked nature of randomization to our estimates by block-bootstrapping standard errors for these estimates. I find suggestively negative point estimates in two of the time slots, but no results are statistically significant.

The substantial noncompliance observed in the data is suggestive of several impor-

tant considerations for randomized trial design in similar settings. It is apparent that instructor fixed effects can be substantial – in our setting, there was stark contrast in the level of experience of the traditional vis-a-vis the active learning instructors in precisely the two sections that attracted the most non-compliers (i.e., this level of experience is also accompanied by renown among students). Absent randomizing instructors to pedagogical modes or restricting the ability of students to change sections (both impossible, practically speaking), the solution to overcoming this threat to identification is replication – namely, to continue to monitor the performance of students in the two instructional modes. In the longer run, sample size increase and balancing of fixed effects as instructors accustom to the new method and students gain comfort with how to perform in such a class stand to strengthen our understanding of the impacts of active learning for student achievement.

Our experience with treatment infidelity in the pilot round of our collaborative study with the City of Philadelphia led us to design a more ambitious and tightly-controlled second iteration of our real estate tax experiment, which is the topic of Chapter 3[4]. Property taxation is the primary tax for most U.S. cities. In fiscal year 2013, 30 percent of all local government revenues and over 73 percent of local taxes came from the property tax Barnett and Vidal [2013]. Yet collection of the tax has, in many cities, been problematic. While some U.S. cities do an excellent job in collecting the tax, receiving over 95 percent of assessed revenues in the year the tax is

[4]This chapter is a co-authored work, with Charles Loeffler, John MacDonald, Holger Sieg, and Robert Inman

due, other cities have over the last ten years done significantly worse – notably Flint (78%), Cleveland (84%), Pittsburgh (86%), Milwaukee (87%), Philadelphia (88%), Detroit (89%), and St. Louis (89%).[5] While Flint, Detroit, Cleveland and Milwaukee are relatively poor cities, Philadelphia and Pittsburgh are not. Among the list of cities with outstanding tax collection records are Buffalo, Birmingham, Houston, and New Orleans. While city poverty is important, it cannot be the whole explanation for low rates of collection. Poor tax administration is likely to be an important contributing factor as well.

This failure to collect the property tax on time creates budget uncertainty at best and budget deficits at worst. Yet collecting the property tax should be straightforward. In contrast to collecting self-reported taxes such as those on income, profits, and sales, property tax obligations equal to the city's assigned assessed value of the taxed property times the city chosen tax rate are known by both the city and the taxpayer. There is no uncertainty as to what is due, or when.[6] Payment is primarily a matter of enforcement. The most common enforcement strategy is the economic stick: fines and penalties. Failure to pay property taxes in time leads to interest penalties sufficiently large that there is no arbitrage advantage to waiting, and perhaps to a significant late fine as well.

[5]For more details, see Chirico et al. [2016].

[6]Much of the current literature on tax compliance has focused on taxpayers truthful reporting of income or sales under the threat of a tax audit; see Slemrod [2007] for a review and more recently the research of Kleven et al. [2011] and Pomeranz [2015].

When a delinquent taxpayer does not respond to penalties and fines, the city can take out a tax lien on the property. A lien does not impose any immediate direct, tangible costs on a taxpayer since payments are typically only realized at the time of a transaction.[7] However, obtaining a tax lien enables the owner of the lien to eventually start a foreclosure process. When the owner of a property located in a city fails to make a payment arrangement on municipal tax levied on his or her property, that property may be sold at auction to allow the city to collect on that unpaid debt. However, the foreclosure process is costly and time intensive.[8] While there are some problems with the effectiveness of the existing enforcement mechanisms, it is only possible to avoid payment by abandoning the property in the long run. Needless to say, this is a very costly option for most owners.

Despite the fact that there are no obvious financial gains to not paying property taxes, we observe that a significant fraction of tax payers do not pay on time. To explain the behavior of these procrastinators, researchers have started to explore the effectiveness of softer, nudge approaches or notification strategies to reinforce the different motivations of tax compliance. This chapter uses a field experiment involving over 19,000 delinquent Philadelphia taxpayers to examine the effectiveness of seven

[7] A city can also sell tax liens to investors to speed up the revenue collection process. Liens often sell at above par prices because of the foreclosure option. But selling liens to "vulture investors" can be politically costly for a city administration.

[8] Auctions are administered in Philadelphia by the Office of the Sheriff. This process of offloading a property at Sheriff's Sale can take nine months to a year.

alternative strategies for improving city property tax collection. Each involves a randomly assigned tax "nudge" of a tardy taxpayer. The first is a simple reminder that the payment is late. The next two involve the reminder plus a threat of a significant sanction if payment is not received by the end of the calendar year: a lien on the home when sold equal to taxes due plus accrued interest and penalties *or* the lien coupled with an immediate sheriff's sale of the home to collect the lien. The final four nudges include the reminder coupled with an appeal to what the tax compliance literature has called a "tax morale" motive for paying one's taxes.[9] The four morale motives included here are: first, a reminder that taxes pay for neighborhood services such as street repairs, trash pick-up and the local park; second, a reminder that taxes pay for important city-wide services such as police protection and public schools; third, a reminder that 9 out of 10 Philadelphians have paid their taxes and you have not; and fourth, a reminder that paying one's taxes is an important obligation of citizenship in a democracy. Tax compliance after receiving one of the seven "nudges"

[9]See Luttmer and Singhal [2014] for a review of the tax morale strategies for tax compliance. They identify three tax morale motivations in the literature, each grounded in a positive gain in utility from the act of paying one's taxes. These include: 1) a motive from reciprocity where the taxpayer recognizes they are part of a larger group playing a non-cooperative game with other taxpayers for the provision of public goods; 2) a motive from peer behavior where the taxpayer gains utility from knowledge that they are part of larger group of contributors; and 3) an intrinsic motivation that provides a direct utility benefit from the act of paying one's taxes. Luttmer and Singhal [2014] also mention taxpayer culture and taxpayer behavior other than utility maximization as additional explanations for the rate of taxpayer compliance.

is then compared to compliance for those who have not received a "nudge" dues to random assignment to a holdout sample.

To understand the potential influence of each nudge, we model tax delinquency as a problem of taxpayer procrastination following Akerlof [1991] and O'Donoghue and Rabin [1999]. Procrastination occurs because of present bias as in O'Donoghue and Rabin [1999] and declining saliency as in Akerlof [1991]. Present bias is always present. Saliency can be nudged by a reminder letter. The reminder letters stressing liens or liens plus the sale of one's home add a future expected cost to non-payment as in the tax compliance model of Allingham and Sandmo [1972]. The reminder letters stressing the tax morale are modeled as utility gains to the procrastinator from paying ones taxes. Economic theory, therefore, plays a central role in the design and implementation of our field experiment. Late payments arise in our model due to lack of salience, lack of deterrence or lack of tax morale. We have designed the field experiment to test the importance of these three competing theories. We show that the treatment effects that are identified by our experiment have a clear interpretation in the context of the parameters of our model. Our experiment is, therefore, designed to explicitly test competing models of behavior as recommended by Levitt and List [2007] and Card et al. [2011].

Our work here is closely related to the recent work of Hallsworth et al. [2014] studying the effect of taxpayer nudges on the timeliness of income tax payments in the UK and to Castro & Scartascini's [2015] study of local property tax payments

in Argentina. Like our study, the amount owed to the tax authorities in these two studies is known by the authority and the taxpayer with certainty; the only issue is payment. As here, the empirical analysis of Hallsworth et al. [2014] follows from a model of taxpayer procrastination. The primary focus of their field experiments is the *framing* of the morale nudge, comparing the effectiveness of what they call a *descriptive* message ("a majority of citizens pay their taxes") to that of an *injunctive* message ("you *should* pay your taxes because"). Our analysis also includes a descriptive message ("9 out 10 taxpayers have paid their tax") and an injunctive message (paying one's taxes is a duty of citizenship"). We differ from the Hallsworth et al. [2014], by including a more strongly worded message on the penalties for non-compliance and by allowing a longer period of study for compliance behavior (3 weeks vs. 6 months in our study). The longer period allows a sharper identification of the saliency of each nudge. Finally, they study compliance for the payment of an important national tax; we study compliance for an important local tax. Like the work here, Castro and Scartascini [2015] study citizen payment of their local property taxes. They also examine effectiveness of separate nudges that stress legal and financial consequences of non-payment, the advantage of payment for the provision of neighborhood services (street lighting), and the fact that seven of ten taxpayers do pay their bills on time. However, they do not consider the effects of saliency nudges independent of the content of the nudges, which is one of the key objectives of our analysis.[10]

[10]We conducted an earlier pilot study of property tax compliance in Philadelphia. The results are reported in Chirico et al. 2016. In contrast to our results here, we find evidence that tax morale

Our experiment supports three central conclusions. First, saliency of the tax obligation matters. A simple reminder letter has both a statistically significant and quantitatively significant impact on the rate of taxpayer compliance, though the effect wears off over time. Second, beyond the simple reminder, the content of the "nudge" matters as well with those stressing rising financial penalties having the greatest impact on compliance. Those appealing to a tax morale – neighborhood, community, peer behavior, and civic duty – were no more successful than the simple reminder letter in inducing additional tax compliance. Third, the marginal revenue benefit of our most effective message is significant, raising $36 in new revenue for each $1 of administrative costs. That said, however, the aggregate effect of nudges on uncollected revenue is modest, bringing in only 5% of all revenue still owed, at least in Philadelphia.

motives drive by public good provision, peer effects, and civic duty can positively impact property tax payment compliance; see footnote 56 below.

Chapter 2

Teacher Turnover in Wisconsin

Literature Review

Because the potential policy implications of turnover in the teaching profession (from human capital and equity/distributional perspectives both) are far-reaching and polypartisan, the literature on turnover-related topics in education is extensive. As relates to this chapter, there are five broad (and often overlapping) categories of inquiry: the relationship between turnover and wages, which has tended to focus on "opportunity wages" outside of the field of education; the relationship between turnover, school demographics, and other nonpecuniary benefits, which has tended to focus on distributional inequalities–whether teachers with certain characteristics are more or less likely to be teaching certain disadvantaged groups; the relationship between turnover and teacher quality as measured by student performance, usually value added (VA); collective bargaining agreements in education, focusing by and large on the implica-

tions (or lack thereof) of seniority-preferential clauses; and the recent phenomenon of specific retention incentives, the provisioning of wage bonuses to teachers willing to teach in high-needs schools.

One of the earliest papers attempting to rigorously investigate turnover was a panel study of teachers in Michigan by Murnane and Olsen (1990), who used college degree field wages outside of education as opportunity wages, finding the expected lower exit rate for teachers with higher wages in teaching relative to the authors' defined alternative. Dolton and Van der Klaauw (1999) use panel data on university graduates in the United Kingdom to estimate a competing risks model of the decision to leave teaching entirely, finding results in line with Murnane and Olsen (1990). Returning to panel studies in the US, Loeb and Page (2000) use PUMS data to get an idea of teacher relative wages in many states and find that dropout rates fall when teacher relative wages are high. Stinebrickner (2002) also uses panel data (this time NLS-72) to track both teachers and non-teachers, focusing in particular on young teachers who leave the profession for long stints, and finds that the best predictor of female exit is recent childbearing, which is an important consideration for all work related to teacher turnover because such a high percentage (76 nationwide) of teachers are female. Lastly, Hanushek, Kain, and Rivkin (2004) focuses on teachers in Texas and emphasizes that the characteristics of students are much stronger factors in predicting teacher exit than are wages (while also affirming the statistical significance of pay).

While wages have been found consistently to have some measurable effect on teacher turnover, it is impossible to explain within-district migration (which constitutes a large portion of switching–as much as 50%) through wage-only channels because contracts are fixed at the district level. As such, another strand of literature has chosen to focus on the nonpecuniary aspects of the decision to take a teaching job–school environment/rapport, student enthusiasm, neighborhood characteristics, etc.–usually by directing attention to a single district so that any wage-based considerations are stifled, as is the case for Boyd et al. (2005) and Engel, Jacob, and Curran (2014). Boyd et al. (2005) track early-career teachers in New York City as they quit or transfer out of the city, and most importantly finds that commuting time is an important, often overlooked aspect of location preference. Engel, Jacob, and Curran (2014) leverages a unique data set from Chicago Public School job fairs which affords them a rather strong measure of teachers' demand for vacancies, neutralizing the influence of school administration's behavior on turnover (through poor match selection or other means). The authors contribute evidence that the school's neighborhood (perhaps due to ambient crime or other reputational effects good and bad) is a better predictor of teachers' preference than distance from home, going somewhat against the grain of Boyd et al. (2005). Scafidi, Sjoquist, and Stinebrickner (2007) examine statewide data from Georgia, but ignore wage effects, choosing instead to focus on disentangling the contributions of low student achievement and minority status to turnover; they find that minority status is the more salient associate of teacher

exit.

The key element missing from all of the above studies is perhaps the most important consideration in the issue of teacher turnover–teacher quality. None of the studies above have student-teacher matched data, and so are unable to directly associate student outcomes with any given teacher. If, with respect to any measure of quality you would like, I find that transitioning teachers are identical to their replacements, the issue of teacher turnover is not, in fact, much of an issue. Thus, the recent trend in the literature to incorporate measures of teacher quality (in large part made possible by a trend towards administrative records allowing students to be linked to teachers and tracked over time) in considerations of teacher turnover has made big strides in addressing the most policy-relevant questions to be asked. The most common and widely accepted measure of teacher quality is VA[11] (in its various guises), and the literature has begun to incorporate such measures into studies of teacher turnover. Hanushek and Rivkin (2010) consider VA as a measure of teacher productivity, and ask if common results of labor search theory (namely that turnover falls with tenure and that turnover is negatively associated with match-specific productivity) continue to hold in the education labor market. In fact, the authors find that the teachers most likely to switch schools are those with low measured match quality, and especially that

[11]The most commonly cited expositions on value-added, its validity, and so on are probably Rivkin, Hanushek, and Kain (2005), an extensive exploration of the predictive powers of empirical Bayes VA measures; and Chetty, Friedman, and Rockoff (2014a) and Chetty, Friedman, and Rockoff (2014b), the largest-scale study of long-term inferences based on VA.

those who leave teaching entirely are those with the lowest match quality. The results are more pronounced for schools with high proportions of low-SES students, which has strong policy implications, as it appears the best teachers in high needs schools are the least likely to change jobs. Goldhaber, Gross, and Player (2007) performs a similar analysis with the longitudinal data of North Carolina and comes to similar conclusions, strengthening the robustness of the results. Lastly, Goldhaber, Lavery, and Theobald (2015) examine the inequity in the distribution of teacher quality by high-needs groups in Washington state, and find that for all three measures of quality (teacher experience, licensure exam score, and VA), the distribution of teachers favors the less needy (as measured by free/reduced-price lunch status, minority status, and low prior academic achievement).

The aforementioned papers have tended to keep the collective bargaining aspect of salary determination for teachers out of the spotlight, if largely for reasons of data restrictions. Nevertheless, it stands to reason to believe that the rigid structure of union-negotiated contracts could serve to contribute in a large way to teacher turnover. Ballou and Podgursky (2002) give much descriptive evidence of the shape of the wage-tenure profile, rooted in a data set collected by the Department of Defense and published by the AFT. They find that seniority premia in education largely mirror those in more traditional white collar professins, that steeper profiles are associated with less turnover, and that district financial and demographic conditions alone are insufficient to explain variation in contracts. Another common (and recently quite

controversial, as evidenced by the contention in the ongoing contract negotiations in Philadelphia) feature of union-negotiated teacher contracts are seniority priviliges–preferential treatments granted to teachers in voluntary and involuntary transfers. Moe (2006) codes contracts from 158 districts in California according to the strength of seniority rights therein guaranteed to teachers and finds that such rights are associated with the distribution of teachers across schools (measuring quality as experience and certification) in a way that serves to harm minorities. Revisiting California with a slightly different sample and definition of the "determinacy" of the contracts with respect to seniority, Koski and Horng (2007) come to the opposite conclusion–that there is no such relationship. As a rebuttal, Anzia and Moe (2014) pin the difference in results on the exclusion in Moe (2006) of small school districts, where it appears that the entrenchment of bureaucracy falters and the rigidity of contract language wane, a claim which they support by repeating their analysis with the inclusion of an interaction for district size–indeed, for small districts the result of Koski and Horng (2007) holds, while the insight of Moe (2006) holds in larger districts. Cohen-Vogel, Feng, and Osborne-Lampkin (2013) use data from Florida and their results align with those of Koski and Horng (2007) (though they neglect to nuance their results by district size).

Finally, an emerging strand of literature is looking at the potential for transfer bonuses and retention incentives to positively affect student outcomes. Fulbeck (2014) analyzes a scheme in place in Denver whereby teachers who choose to trans-

fer to high-needs schools (low-performing) are given recurring bonus pay, and those initially stationed there are given retention incentives. She concludes that recipients of incentives are significantly less likely to switch jobs, as driven by a reduction in district exit rates and especially by teachers whos incentive payments exceed $5,000. Glazerman et al. (2013) evaluate the Talent Transfer Initiative, a randomized controlled trial conducted in 10 districts whereby high-performance teachers were given $20,000 over the course of two years as reward for transferring the identified high-needs schools, and conclude that there were significant effects on teacher retention as well as on student outcomes.

Two highly germane papers investigate the impact in Wisconsin on teachers of Governor Scott Walker's flagship policy, Act 10, which severely limited the scope for collective bargaining in the state. Litten (2016) uses differences in contract renewal dates surrounding the policy's enactment to evince the effect of unionization on teachers' wages, and finds the lack of union bargaining power reduced teacher compensation by 8%. Biasi (2017) constructs value-added measures from grade-level test results and concludes that the move to individually-negotiated salaries in some districts had a significant impact on teacher quality and student outcomes in such districts, while also cautioning that most of these gains are competition-based, so that scaling up the system state-wide would have an impact limited to a boost from the exit of low-quality teachers.

Data

The State of Wisconsin's Department of Public Instruction (DPI) releases annual Salary, Position & Demographic reports through the WISEstaff data collection system. These reports represent "a point-in-time collection of all staff members in public schools as of the 3rd Friday of September..." (Public Instruction 2017a), and will serve as the primary source of data on teachers in this chapter. Data are available at the position-teacher level cross-sectionally, with each entry in a given year corresponding to one of possibly several positions/assignments held by each school district employee[12]. Identifiers in each file permit unique identification of an employee within a given year, but this identifier does not follow teachers between years[13]. To overcome this substantial hurdle to identifying teacher mobility, data are first fed through the matching algorithm described in further detail in the Appendix. Essentially, I are aided by the availability of various imperfect identifiers which should be more stable over time, most crucially teachers' first and last names and year of birth. By building on these covariates and incorporating some limited fuzzy matching techniques, I construct a panel of teachers spanning the 1994-95 academic year (AY) through

[12]Many teachers (and other district employees) serve in multiple roles within a school/district, for example as a coach, part-time program aide, or department head. Each of these is filed as a separate observation in the DPI system, though salary information is given at the teacher as opposed to the assignment level.

[13]From AY2011-12, a field called the File Number appears to allow longitudinal tracking of teachers. I use this in part to validate the matching algorithm; see the Appendix.

AY2015-16[14] consisting of 3,588,614 teacher-position-year observations. The matching algorithm necessitates elimination of 26,304 (0.7%) observations over all 21 years on account of belonging to teachers who could not be uniquely identified in a given year of data due to exact overlap of their first name, last name, and birth year fields with another teacher in the data[15].

Specific to the exercise at hand, with data reliability and precision in mind, I make the following series of further restrictions on the data. The introduction of Wisconsin Act 10 introduced a substantial structural break in the labor market for Wisconsin teachers, so I include only data from 2000-2010 to avoid conflating the effects of this policy on teacher turnover with the earlier functioning of the labor market (i.e., I do not want to mix the results from distinct equilibria of the teacher labor market, but would instead prefer to analyze the pre- and post-Act-10 markets separately). I drop all employees who are not full-time, full-year regular teachers of a major core subject (all-purpose elementary teachers or English/Math) at a single regular public school with a Bachelor's or Master's degree and fewer than 35 years' recorded experience; taken together, these restrictions eliminate 79% of employees, the lion's share of which come from eliminating substitutes/support staff and teachers of on-core subjects[16]. I then eliminate teachers with missing information on their

[14]For brevity, I herein refer to academic years by the spring year, e.g., AY2003-04 will be simply 2004.

[15]Technically, I use a slightly modified version of the name strings in making these eliminations which, for example, eliminates initials – see Appendix.

[16]I also eliminate any teacher who appears in any role besides "Teacher" in any year. In particular,

subsequent school or district and teachers with instability in their recorded ethnicity, as well as teachers not categorized as white, black, or Hispanic, eliminating a further 0.2% of all employees[17]. Finally, I drop teachers' multiple positions by keeping only the highest-intensity position for each teacher, as measured by full-time equivalency, resulting in a final count of 282,797 teacher-year observations – 49,325s in 449 districts and 2,296.

The data used for the incorporation of counterfactual salary calculations is largely the same, but with a few noteworthy differences. First, as noted in Footnote 16, the main turnover data eliminated some teachers who transitioned in and out of being categorized as a full-time teacher due to the muddling effects thereof on defining turnover. This concern not being relevant to constructing the salary schedules, it is not imposed for this data. Next, because all regular teachers are covered by the

this eliminates a nontrivial number of educators who begin their career with an "ease-in" period, take a mid-career "leave" speckled with a transition to substitute teaching – perhaps during their child's infancy – or end it with a "soft retirement" period, during which they act as a substitute teacher at some point in the midst of a career otherwise focused on teaching. Such teachers often have part-time roles at several local schools, which introduces sufficient ambiguity in the definition of mobility so as to obscure interpretation of results, so I opt for a stricter definition of full-time teaching than is completely necessary.

[17]Wisconsin teachers are predominantly white (96%). As noted in the Appendix, I also use the panel data to correct noise found in recorded ethnicity and gender over time for some teachers. In the final sample, 427 and 345 teachers had their ethnicity and gender (respectively) adjusted in some year(s).

same collective bargaining agreement, the salary imputation data is less restrictive with respect to the subject codes excluded from the data, and generally includes any teacher not in special education. This data also ignores instability in recorded gender and ethnicity within a teacher.

Finally, the salary data loses observations that are present in the turnover data based on a series of cuts which are either required for COBS to function, or else substantially increase the reliability of its output. The most noteworthy/far-reaching of these numerical restrictions is to eliminate any teachers working in districts where there are not at least 20 total teachers in each degree track for that year. While ultimately arbitrary, this number is reasonable to limit the potential effect of an individual teacher on an exercise determining 35 levels of pay with minimal functional form restrictions. The other numerical flags require both the BA & MA track to be represented at a district, for at least 7 distinct levels of experience to be represented within a degree track, and for at least 5 unique values of the two measures of pay (salary and fringe benefits) to be available in each degree track; all teachers at districts failing at least one of these tests is dropped.

The sum total of all of these restrictions leaves us with an analysis sample of 356,265 teacher-year observations to be used to estimate pay scales, made up of 65,069 individual teachers in 209 districts over 11 years. In total, there is sufficient data to fit 3,708 $y_t(\tau, c, d)$ curves, an average of roughly 100 observations per curve. Ultimately around 22% of teachers have missing salary information[18], mostly in rural districts

[18]HKR include like-minded restrictions, but combine teachers of different certification within an

or other districts with only one or two schools and a small number of students.

I supplement the WISEstaff data set in several ways to incorporate information about other characteristics of schools and districts in Wisconsin. To get school- and district-level measures of socioeconomic makeup (percentage of students who are black or Hispanic or eligible for free/reduced lunches) and community type/urbanicity, I tap the Universe Surveys from the National Center for Education Statistics' Common Core of Data, which provide this information on a yearly basis for all years in the study[19,20]. At the district level, I also use this data to compute class size and the size of the student body.

Lastly, I turn to DPI's public data again to get school- and district-level performance metrics. While Hanushek, Kain, and Rivkin (2004) were able to obtain school- and district-level average scale scores on a standardized test in Texas, such a metric is not publicly available in Wisconsin for all years. Instead, I calculate student pro-

experience level, despite the headline importance of this factor to teacher pay – median pay at a given level of experience is on average 17 higher for those with a Master's degree.

[19]The method of recording urbanicity by the Common Core switched from being "metropolitan-centric" to being "urban-centric" for Wisconsin from 2006 (Sable 2009). I map the corresponding codes to match those used by HKR as well as possible, and use the data file from 2006, which has both types of code for all US districts, to confirm that the pre- and post-2006 correspondence is by-and-large working as intended. For a small number of districts/schools with missing urbanicity codes in certain years, I use information about that entity from other years to inform urbanicity.

[20]Further, the WKCE data does not include a standard deviation field even in those years when the school average scale score is available, precluding any attempt to standardize test scores and put the data here on equal footing with that of HKR.

ficiency rates for each school and district as the percentage of test-takers deemed to be at grade level in mathematics or reading in a given year on the Wisconsin Knowledge and Concepts Examination (WKCE), which is administered to 4th, 8th, and 10th-grade students.

Salary Scale Imputation with Constrained B-Splines

For many years, the ubiquitous characteristic of collectively bargained teachers' contracts has been the salary table, which gives a mapping from the calendar year, a teacher's experience (their length of tenure at the current district), and their certification (typically Master's vs. Bachelor's degree) to their wage. This table gives current teachers a clear understanding of how their pay will advance as a function of their labor inputs, and thereby gives forward-looking potential teachers and potential migrant teachers a clear understanding of their would-be pay arcs under a district-switching decision-making framework, especially given that this information is typically openly available.

It would behoove an econometrician seeking to understand education labor market dynamics, then, to incorporate this information on future pay into their statistical modeling framework. Unfortunately, this data is typically not available in a format lending itself to easy analysis at scale – whether locked inside idiosyncratically formatted and sporadically-available contract PDFs or hidden behind large-scale freedom of information act inquiries, the temporal and financial costs of scraping such data into

a usable form can be substantial.

Much more common in empirical settings is access to teacher-year-level salary data of the form $y_{i,t} = y(\tau_{i,t}, c_{i,t}, d_{i,t}) + \varepsilon_{i,t}$, where $\tau_{i,t}$, $c_{i,t}$ and $d_{i,t}$ are the tenure, certification, and district of teacher i in year t, and $\varepsilon_{i,t}$ represents unaccounted factors affecting the wage (e.g., not all teachers work full time, some teachers split their time among duties yielding different pay levels, and many teachers supplement their income with additional duties like coaching). Here I consider one approach and some empirical lessons for trying to estimate the underlying mapping $y(\tau, c, d)$ from such data.

There are a multitude of inference/imputation techniques suitable to the inference of a latent function of unknown parametric form available in the statistician/econometrician's palette. The powerful flexibility of nonparametric approaches (local regression, splines, Random Fourier Feature expansions) is a double-edged sword; as it happens, in this particular setting, even if I know linearity is not a reasonable functional form restriction, I do know some very basic properties of the underlying tenure-wage curves that will be violated in general by uninformed estimation techniques. In particular, I know that such tenure-wage curves are non-decreasing and that they are non-negative, i.e., $y(\tau', c, d) >= y(\tau, c, d)$ whenever $\tau' >= \tau$, and $y(0, c, d) \geq 0$.

He and Ng (1999) introduce a linear programming approach to incorporating monotonicity, curvature, and pointwise restraints to quantile regression spline esti-

mation techniques, and Ng and Maechler (2007) present an overview of the R package `cobs` which gives an efficient implementation of this approach (COBS standing for Constrained B-Splines; B-splines are computationally-efficient basis functions for degree k splines). The basic idea of quantile regression spline estimation is to swap out the standard squared loss function for a quantile-dependent weighted absolute loss function to target conditional quantiles instead of conditional means. Monotonicity, point, and curvature restrictions enter as penalized terms to the objective function; `cobs` expresses this in a fashion which facilitates the application of standard linear programming techniques for efficiency, and handles internally the issues of knot selection and penalty parameter assignment through cross-validation.

I implement and fine-tune this general approach with an eye to being as minimally-invasive as possible. The first innovation is required by the poor performance of standard COBS fit in extrapolation. Data sparsity in smaller districts means that it is often the case that only a small range of τ values are observed in a given year-certification-district. Monotonicty constraints are only built into the B-spline routine internally; the underlying basis functions may produce decreasing fits outside the observed range of data. To overcome this, I take a cue from the literature tackling Runge's Phenomenon (Runge 1901), wherein polynomial approximations tend to exhibit extreme oscillations in extrapolation. This issue is one of the motivations behind natural cubic/smoothing splines (see, e.g., Friedman, Hastie, and Tibshirani 2001; Wahba 1990; Green and Silverman 1993; or de Boor 1978), which handle this is-

sue by using a simple linear basis function outside the outermost interpolating knots. I incorporate this technique of linear extension only when necessary by testing the COBS fit for monotonicity; τ values failing this constraint are replaced by extending the final non-decreasing fit values through the end of the range of extrapolation.

Next, a major shortcoming of COBS for this context is its limit to one-dimensional spline fits; while techniques for nonparametric B-spline fits are available in arbitrary dimensions (see de Boor 1978), at present COBS is only capable of imposing monotonicity on one dimension of a curve. In my context, however, $y(\tau, c)$ is increasing not only with respect to τ, but also with respect to c (as, without fail until only very recently in Wisconsin, certification was rewarded with a Master's premium, typically a percentage increase in wage). One solution would be to generalize the implementation of COBS to handle a second dimension by simply adding penalty terms along this dimesion[21]. I abandon this approach because of the categorical nature of the certification dimension – there are not numerical units to the difference between having a Master's vs. Bachelor's degree. The assignment of such a number required by this approach would itself become an implementation hyperparameter, meaning that the ultimate fit would itself be sensitive to the particular choice of continuous representation.

[21]Not to mention the empirical reality that the two-lane dichotomy is in fact false – it is very common, nearly ubiquitous, for contracts to offer separate lanes for teachers with the same completed certification, but different levels of progress towards completing further certification. As this dimension is impossible to glean from my data, I exclude it from the imputation exercise.

Instead, I use a two-step procedure to fit the Bachelor's and Master's pay tracks in serial. In the first step, I fit the Bachelor's career track as a typical one-dimensional COBS fit. In the second step, I first construct Master's premia for each observation by subtracting out the predicted Bachelor's pay corresponding to each observed level of tenure for a teacher with a Master's degree. I then use COBS to fit a non-decreasing Master's premium curve over all tenure levels on these residuals, before finally adding the Master's premium and Bachelor's fit curves to get the overall Master's fit curve. Monotonicity of the result is guaranteed by forcing upwards monotonicity on the Master's premium, a restriction in line with the empirical observation that Master's degree pay is often simply a fixed-percentage rise over the corresponding Bachelor's pay.

My implementation was also aided by the imposition of weak concavity on the $y(\tau, BA)$. While not a theoretically-assured functional form restriction[22], concavity improves the goodness of fit notably. Small-sample district-level observations and simple reduced-form regressions of wages versus quadratic forms in experience support this shape's validity. A variety of contracts obtained from a database for teachers in nearby Michigan also meet this condition, and the decrease in marginal returns to

[22]In fact, in reality tenure-wage curves are often piecewise convex – year-over-year rises are specified as a percentage bump which eventually levels off to either linear increase or maxes out and flattens. Nevertheless, the degree of convexity in that section of the curve tends to be low, which leads COBS to fit a good linear approximation there. The lack of a concavity restriction on the Master's pay track allows fit curves which fit this pattern for this lane.

experience is also commonly found in the wider study of labor markets[23]. I do not impose this restriction on the fit for the Master's premium (the only restrictions there being non-negativity at 0 and, as mentioned, an increasing relationship with tenure).

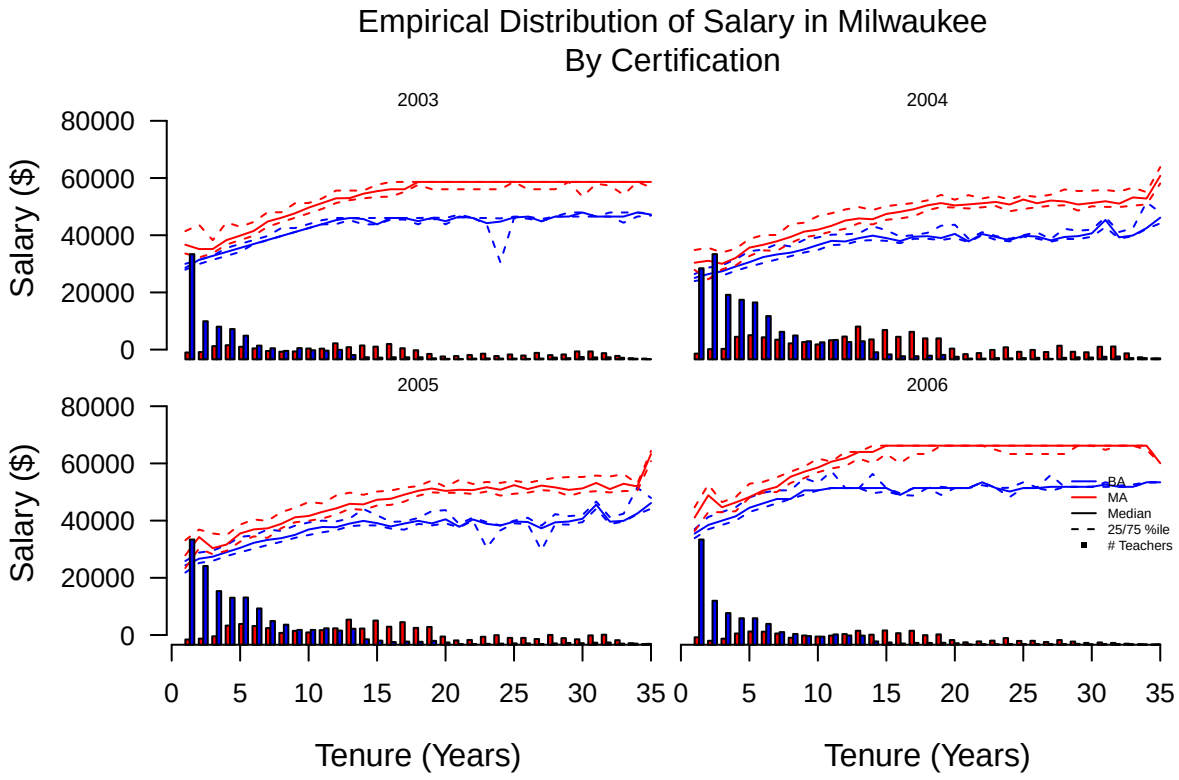

Figure 2.1: Pay in Milwaukee, 2003-2006

As an illustrative example of the patterns in the data I wish to quantify and formalize, I turn briefly now to Milwaukee Public Schools, the largest district in Wisconsin with roughly 32,387.73 teachers per year. Figure 2.1 depicts key moments of the empirical distribution of salary in 4 years at Milwaukee Public Schools, broken down by tenure and certification. The central lines on each plot (Bachelor's pay

[23]See, e.g., Heckman, Lochner, and Todd (2003).

track in blue, Master's pay track in red) are the empirical median levels of pay, and thus give a rough approximation to $y(\tau, c)$. The dashed-line intervals on either side represent the 25th and 75th percentiles.

Notably, these intervals and the medians themselves tend to get quite noisy at later stages in the career, especially for the Bachelor's track. This fact that reflects the almost universal certification of teachers by about 15 years into their career. This is reflected in the bar graph below each set of curves, which shows the distribution of teachers in each certification track by tenure. Almost all new teachers start with only a Bachelor's degree; the relative presence of Master's degrees grows over time as more teachers certify mid-career.

I can also note two more key empirical facts from this plot. First, the vanishing presence of teachers in both certification tracks leads the empirical median to be a poor approximation of $y(\tau, c)$ since it frequently fails to respect the fundamental monotonicity constraint discussed above. With respect to tenure, this tends to affect the Bachelor's track later in the career as more teachers certify, and the Master's track very early in the career before teachers certify. The monotonicity with respect to c of the median wage is mostly maintained here for Milwaukee, but this is not always the case; my estimation procedure is thus careful to impose these restrictions internally.

Second, structural breaks are an important empirical phenomenon in this context. Each time a contract is renegotiated at a district, the tenure-wage curves can

potentially change shape dramatically. It is with this in mind that I refrain, given my ignorance with respect to when such structural breaks occur, from combining information from adjacent years in fitting a given year's curve, an approach which would substantially enhance the statistical power available to fit contracts for sparsely-populated districts. Such a structural break is apparent in Milwaukee, for example, between 2003 and 2004 and between 2005 and 2006, where the shape of the Master's pay scale has shifted notably. While I eschew, for example, full Bayesian estimation of structural breaks in a given district, such techniques are applicable and worthy of future exploration.

Goodness of Fit

Returning to the motivating example illustrated in Figure 2.1, I turn first to the performance in Milwaukee, where, given the relatively large sample size, performance is expected to be very good. Indeed this is the case, as seen in Figure 2.2. The COBS fit has retained all the salient features of the empirical median return to experience and certification, while simultaneously improving over this nonparametric conditional median by ironing out nonmonotonicities found empirically as a result of small-sample bias.

Perhaps more telling is the goodness of fit in minimally small districts. Four such examples are featured in Figure 2.3. These four districts just barely satisfy the sample restriction that at least 20 teachers be present in both the BA and MA pay track (each

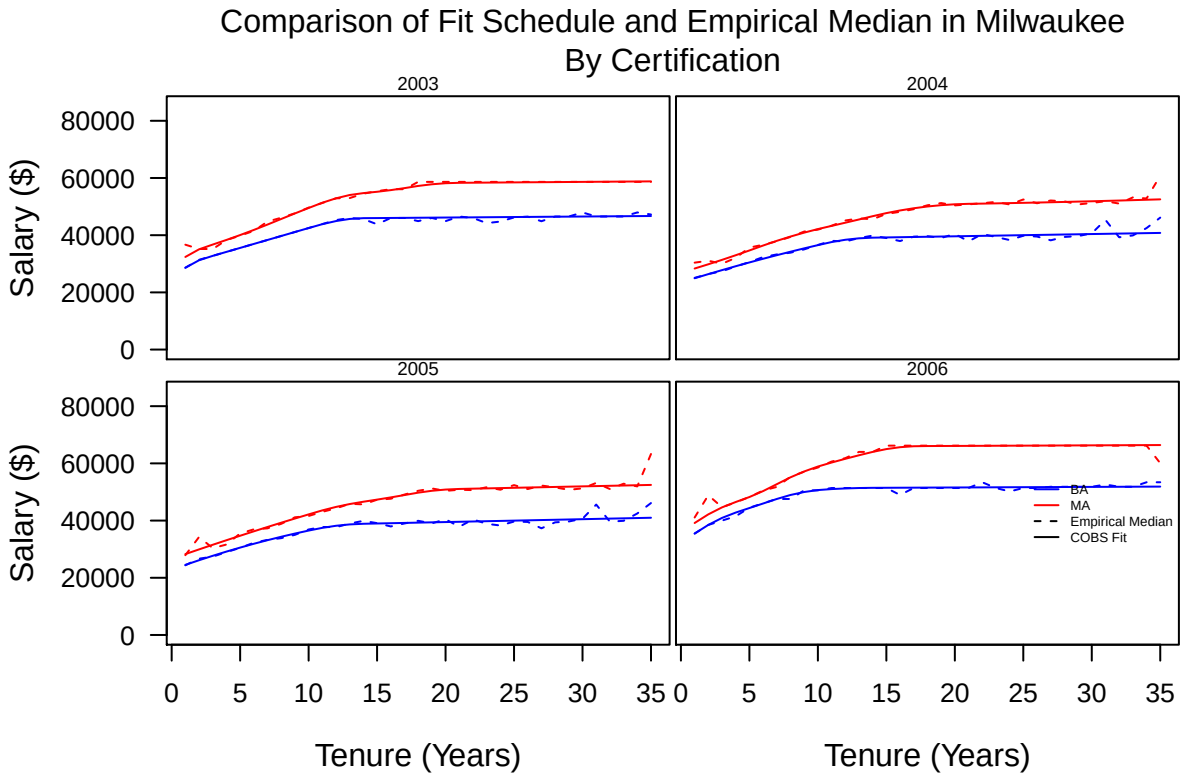

Figure 2.2: Estimation Results for Milwaukee, 2003-2006

has fewer than 42 teachers, and only in a single year); for this reason, rather than plot the empirical median, I simply present the full distribution of wage, experience, and certification in these districts. Here again the COBS fit captures the essence of the wage-tenure curve even in these sparsely-staffed districts. Both Montello and Manawa evince the importance of the non-negativity constraint on extrapolated values of the Master's premium – given the absence of teachers so certified prior to the fifth year of experience, some supplementary discipline is necessary to prevent the tail of this curve from dipping below that for the Bachelor's lane.

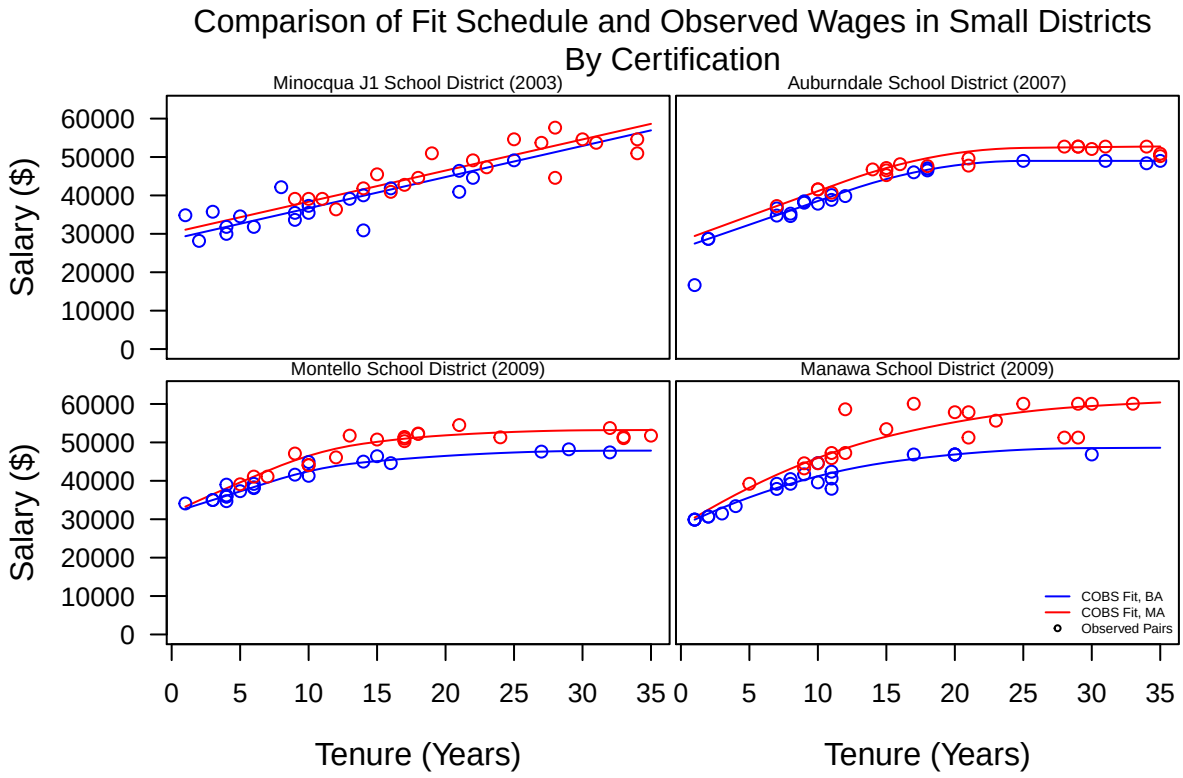

Figure 2.3: Estimation Results for Selected Sparse Districts

A final check on the validity of the imputation procedure would be to compare

fit schedules side-by-side with the true schedules, e.g. through root mean-squared error. As mentioned, this is typically a difficult undertaking on a mass scale since the true schedules may be hard to come by in a parseable electronic format. Usually, a smaller-scale version of this exercise would be possible through sampling, say, 5-10% of districts at random and spending the time to extract actual schedules by hand for this purpose. Unfortunately, with the passage of Act 10 and the abandonment of collective bargaining in many districts, electronic copies of legacy contracts became hard to come by – none of the large districts I contacted (nor their former union representatives) had access to old copies of contracts they were willing to share, nor could I find any but a very small number of these contracts online. I present here the comparison of COBS-produced fit to true schedule in three district-year combinations for which I could actually obtain the true schedule[24].

As seen in Figure 2.4, the resulting fit is generally superb[25]. Only for the Bachelor's track in Monona Grove in 2009 does the COBS-fit curve depart substantially from

[24]These contracts and a few others from outside of the study time frame are available upon request.

[25]In terms of objective measures of the fit, the mean absolute error is $1,762, while the overall median error is $812. This is evidence against the assumption built into the COBS routine of 0-median errors $\varepsilon_{i,t}$, and is understandable – it is not uncommon for teachers to earn supplementary pay from coaching or extra teaching duties that would push them above their salary-schedule-dictated pay grade. With a more complete set of training data, one could potentially account for this by treating the quantile of the data targeted by COBS (.5 by default, i.e., COBS is median-targeted) as a hyperparameter to be fit by cross-validation to prevent overfitting (see Stone 1974 or Friedman, Hastie, and Tibshirani (2001)).

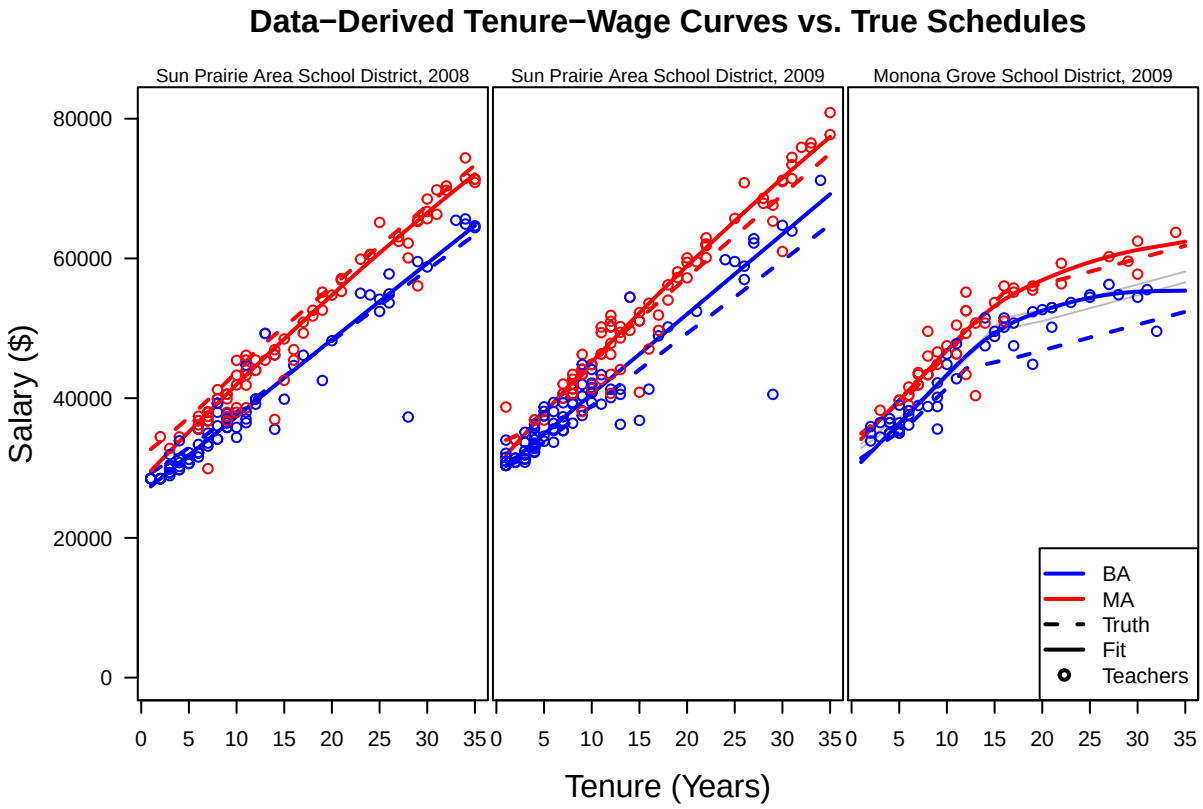

Figure 2.4: Comparison of True Contracted Schedule with Output of Imputation

the true contracted schedule. Moreover, this departure is likely attributable to the oversimplification taken in this chapter of restricting pay to follow only two "lanes" (Bachelor's and Master's degrees), when in reality districts often differentiate among holders of these degrees by rewarding those with more credit-hours of supplementary coursework under their belts in pursuit of continued learning or a higher degree – in fact, such coursework is often required of Bachelor's-certified teachers, which means it is likely that the later-career Bachelor-certified teachers observed in Monona Grove are actually being paid according to a higher lane. This is exactly what is depicted by the gray lines on the Monona Grove plot, which show the BA+12 and BA+24 lanes are more representative of instructors at the later stages of their career in Monona Grove (the data lack any way of detecting a given teacher's extracurricular credit accumulation).

I take the above as strongly affirming the utility of COBS as a tool for constructing wage-tenure curves from teacher-level salary data. It is able to gloss over noise-induced non-monotonicities in the empirical median, not just with the rich data found in urban districts, but also in sparsely-populated districts. Moreover, as explored in the case of Monona Grove School District, COBS can be seen as doing a good job of capturing an aspect of the data which is still latent (namely, the degree of progress towards further certification), and of being closer to the "true" schedules that teachers use to make mobility decisions (since it is likely that teachers are able to anticipate extra income from holding multiple roles and factor this into their as-

sessment of a wage offer). Lastly, the COBS routine is computationally attractive – embarassingly parallelizable and implemented very efficiently, the whole routine runs in a few minutes.

Turnover in Wisconsin

	Percent of Teachers Who				
Teacher Experience	Remain in Same School	Change Schools Within District	Switch Districts	Exit Wisconsin Public Schools	Number of Teachers
1-3 years	79.7	7.1	6.0	7.2	41,042
4-6 years	86.6	5.6	3.3	4.5	37,770
7-11 years	90.6	5.0	1.8	2.6	54,623
12-30 years	92.4	4.1	0.6	2.9	129,002
>30 years	80.3	6.4	1.1	12.2	20,360
All	88.6	5.1	2.0	4.3	282,797

Table 2.1: Year-to-year Transitions of Teachers by Experience, 2000-10

I move now to the core focus of my analysis, examining the distinguishing features of turnover in the teacher labor market in Wisconsin. Table 2.1 replicates Table 1 of Hanushek, Kain, and Rivkin (2004), and as HKR found in Texas, most turnover in

Wisconsin is happening within districts and out of the profession[26,27]. In Wisconsin, the fraction of teachers transitioning among districts is vanishingly small after a "burn-in" period of roughly 6 years – only 1% of such teachers do so (compared with 3.1% for the comparable group in HKR), but is still relatively highest among the youngest teachers – roughly twice as high for the "probationary" teachers (1-3 years' experience) as for teachers with 7-11 years' experience in both states.

By contrast, movement patterns within districts in the two states are very similar, lending weight to teachers "earning their stripes" within a district to be able to choose the best schools as a privilege of seniority. As expected, I also observe a U-shaped pattern in teachers exiting Wisconsin public schools, which jibes with there being two types of quits. Early-career quitters change to private schools, change state of residence, or change professions; late-career quitters retire – especially evident among teachers with more than 30 years' experience, a group which sees a mass exodus of fully 10 percent of its teachers annually. Results not included here break down the

[26]This and subsequent analyses were greatly facilitated by several facilities of the R programming language, for which due credit must be given to R Core Team (2016), RStudio Team (2017), Dowle and Srinivasan (2017), Xie (2016), Leifeld (2013), Dahl (2009), Henningsen and Toomet (2011), Zeileis and Hothorn (2002), Zeileis (2004), Zeileis (2006) and Croissant (2012).

[27]I also note that some "turnover" identified by teachers not appearing at the same school in the following year is in fact spurious – Public Instruction (2011) identifies a number of instances of school districts merging during the timeframe of my analysis and hence disappearing from the data altogether. I take care to reset the district and school switch identifiers off for these 82 teachers if they appear in the newly-formed district in the subsequent period.

exit rates by experience level, where this dichotomy is even more dramatic – first-year exit rates are about 8 percent and quickly level off at around 2 percent before spiking again past around 25 years.

As examined further below, the low rate of switches between districts appears to be owing to the generally more rural nature of Wisconsin vis--vis. Texas. To wit, Milwaukee is the only major urban area in the state, and its population (2010 Census) of 594,833 would rank 7th in Texas. This means that two major types of movers in the HKR data – Large Urban - Large Urban and Suburban - Large Urban – are limited within the state to ending up in a relatively minor metropolitan area. HKR don't provide any results disaggregated by city, precluding any attempts to compare these numbers more comparably to those that would obtain from eliminating the largest cities in Texas.

Moving from the aggregate numbers to begin to examine heterogeneity in turnover, Table 2.2 replicates HKR Table 2, and reverberates its most important conclusions. HKR argue that there is little support for the idea that scores of young teachers are using large urban schools as a training ground before "settling down" with easier assignments in the suburb, based on the general low level of turnover from Large Urban districts. I affirm the scarcity of transitions from districts in Milwaukee, while also noting that such a path is certainly present, as evidenced by the plurality of those who do leave Large Urban districts ending up in a Surburban district in both settings. HKR also observe that the likelihoods of remaining in the same school and of

Origin Community	Percent of Teachers Who Move to				Number Teachers Changing Districts	Percent of Origin Teachers	Change in Share of Teachers 2000-06
	Large Urban	Small Urban	Suburban	Rural			
I. All teachers							
Large Urban	17.4	15.8	48.7	18.1	819	2.7	0.4%
Small Urban	3.7	13.4	44.8	38.1	640	1.2	0.1%
Suburban	3.4	16.0	44.2	36.4	1,408	1.9	3.6%
Rural	0.6	11.2	24.2	63.9	2,794	2.2	-4.1%
II. Probationary teachers (1-3 years experience)							
Large Urban	15.8	17.9	47.6	18.8	437	5.0	
Small Urban	5.1	14.0	46.2	34.8	271	3.9	
Suburban	4.3	16.2	41.0	38.4	561	5.5	
Rural	0.3	10.9	24.4	64.5	1,204	8.0	

Table 2.2: Destination Community Type for Teachers Changing Districts, by Origin Community Type and Teacher Experience Level

quitting are roughly the same for urban and suburban teachers, an observation which I can confirm in Wisconsin. I further note that while Table 2.2 only presents a cross-sectional picture, the career-long trend reaffirms this – only 3.2% of teachers starting their careers at a large urban district ever work at a suburban district. Lastly, I echo the suggestion of HKR that this phenomenon cannot be driven purely by demand-side constraints – in my time period of observation, I observe only 1,459 urban teachers change districts, whereas 3,211 teachers were hired in suburban districts, though of course this does not rule out arguments based for example on stricter screening of applicants transferring from urban districts.

I note, however, that though tales of flight from troubled urban districts are apparently anecdotal, they are far from apocryphal. To wit, while 50 percent of districts have a net inflow (arrivals less departures) of four or fewer teachers (in absolute value), Milwaukee's net outflow was 533 teachers, and the five highest-inflow districts, all suburbs of Milwaukee or districts adjacent the main university town of Madison, saw in total an inflow of 229 teachers in this time. This being a two-sided market, this state of affairs is perhaps largely attributable to the dynamic nature of student populations at these districts – but these, as well, are reflective of the appeal of the districts to parents (and teachers as parents).

As mentioned in the discussion of Table 2.1, the major difference with respect to quantities observed in Texas appears to be driven by differences in the urban landscape between Texas and Wisconsin[28]. This is supported by the overall similarity

[28]I also note a difference (as found in Table 2.2) in the relative shift in population among com-

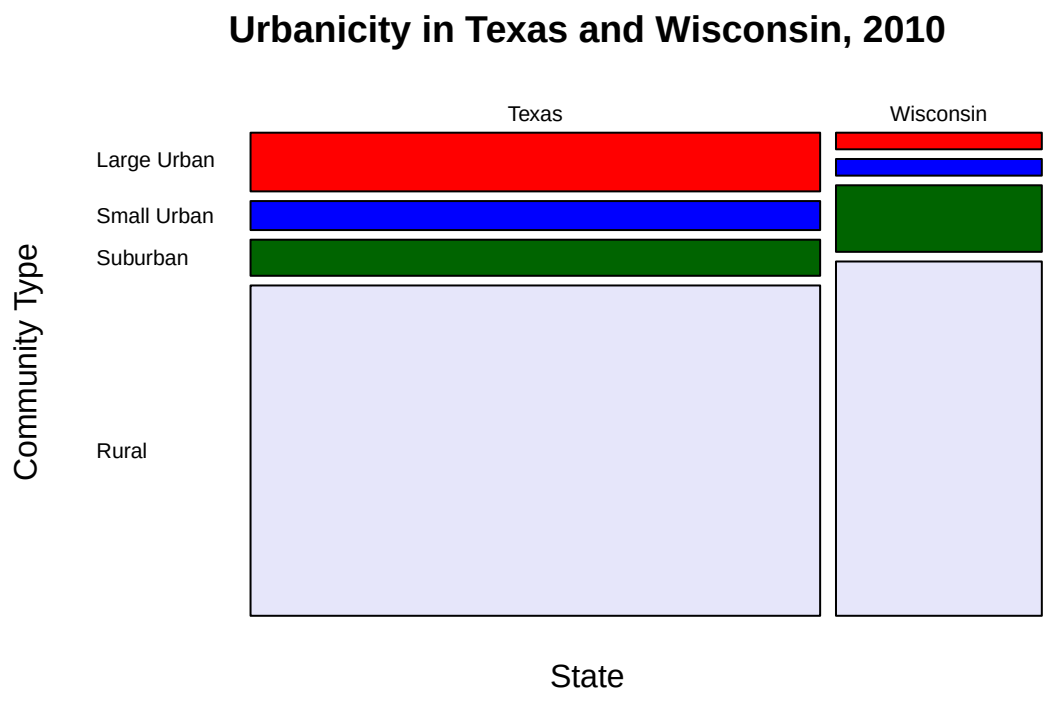

Figure 2.5: Comparison of the Prevalence of Different Community Types

of magnitudes of transition rates to community types besides Large Urban in the two papers. Figure 2.5 depicts this difference in landscape by comparing the distribution of community types in Texas and Wisconsin in 2010 (bar widths reflect the relative quantity of districts in Texas and Wisconsin). While both states are majority-rural, the non-rural part of Texas is comparatively urbanized, whereas more than 90% of Wisconsin districts are non-urban.

Returning to Table 2.2, I see that, as in HKR, the "stickiest" community type is Rural – over 60% of Rural teachers remain Rural in both papers, and even fewer

munity types between the two states – Texas observed dramatic changes in its community type distribution over the period of study of only 4 years, while Wisconsin only saw some movement from Rural to Suburban communities over a longer period of 11 years.

Rural Wisconsin teachers end up in a big city than is the case for Texas. This may reflect the similarity in prevalence of rural districts in the two states and a natural similarity in preferences of rural teachers and districts. Lastly, I also find that the community type transition patterns of younger teachers as compared to all teachers are broadly similar.

Table 2.3 replicates Table 3 of HKR, and again confirms its most important insights. Raw salary differentials predict teacher mobility, but the average pay differential is not on average very large – only about $325, or 1.7% higher than the counterfactually expected wage that would have obtained had the district-switching teacher remained in their current district[29]. This premium increases with age for both male and female teachers.

One potential explanation of the weakness of the wage results is that there simply is not sufficient heterogeneity among available contracts to generate mobility incentives. Figure 2.6 demonstrates that this is not likely the case. No matter their current experience or certification level, a teacher in a district paying the 25th percentile of wages for that experience-certification cell would gain on average 17% by changing to a district at the 75th percentile. Especially for younger teachers, this potential gain would accumulate annually to become a hefty sum over the course of the career –

[29]There are 777 teachers in the data who skipped one or more years before reappearing at different school or district (perhaps representing leaves of absence for retraining or re-adjustment). For such teachers, the counterfactual subsequent experience and reference curves are taken from their next year in the data, rather than from simply incrementing their experience by one.

	Men by Experience Class			Women by Experience Class			All Teachers 0-9 Years
	1-3 years	4-6 years	7-11 years	1-3 years	4-6 years	7-11 years	
Base year salary (log)	-0.001	0.015	0.036	0.001	0.022	0.010	0.009
	(0.009)	(0.011)	(0.015)	(0.005)	(0.007)	(0.010)	(0.003)
Adjusted salary[a] (log)	0.011	0.003	0.024	0.002	0.014	0.015	0.008
	(0.007)	(0.009)	(0.012)	(0.004)	(0.006)	(0.008)	(0.003)
Percent proficient	4.2%	3.0%	2.5%	6.3%	5.7%	5.3%	5.4%
	(0.7%)	(0.8%)	(1.0%)	(0.4%)	(0.5%)	(0.6%)	(0.2%)
Percent Hispanic	-1.4%	-0.3%	-0.2%	-1.7%	-1.6%	-1.1%	-1.4%
	(0.3%)	(0.4%)	(0.5%)	(0.2%)	(0.2%)	(0.3%)	(0.1%)
Percent black	-5.4%	-2.1%	-3.8%	-8.6%	-6.8%	-6.9%	-7.0%
	(1.0%)	(1.1%)	(1.2%)	(0.6%)	(0.7%)	(0.8%)	(0.3%)
Percent subsidized lunch	-7.4%	-3.7%	-4.4%	-9.5%	-7.0%	-7.6%	-7.9%
	(1.1%)	(1.4%)	(1.7%)	(0.6%)	(0.9%)	(1.0%)	(0.4%)

Note: a. Adjusted salary is the residual of log salary by district and experience level on 12 regional indicators, three urbanicity indicators, and the district percentages proficient on the WKCE exam, black, Hispanic, and low income.

Table 2.3: Average Change in Salary and District Student Characteristics (and Standard Deviations) for Teachers Changing Districts, by Gender and Experience

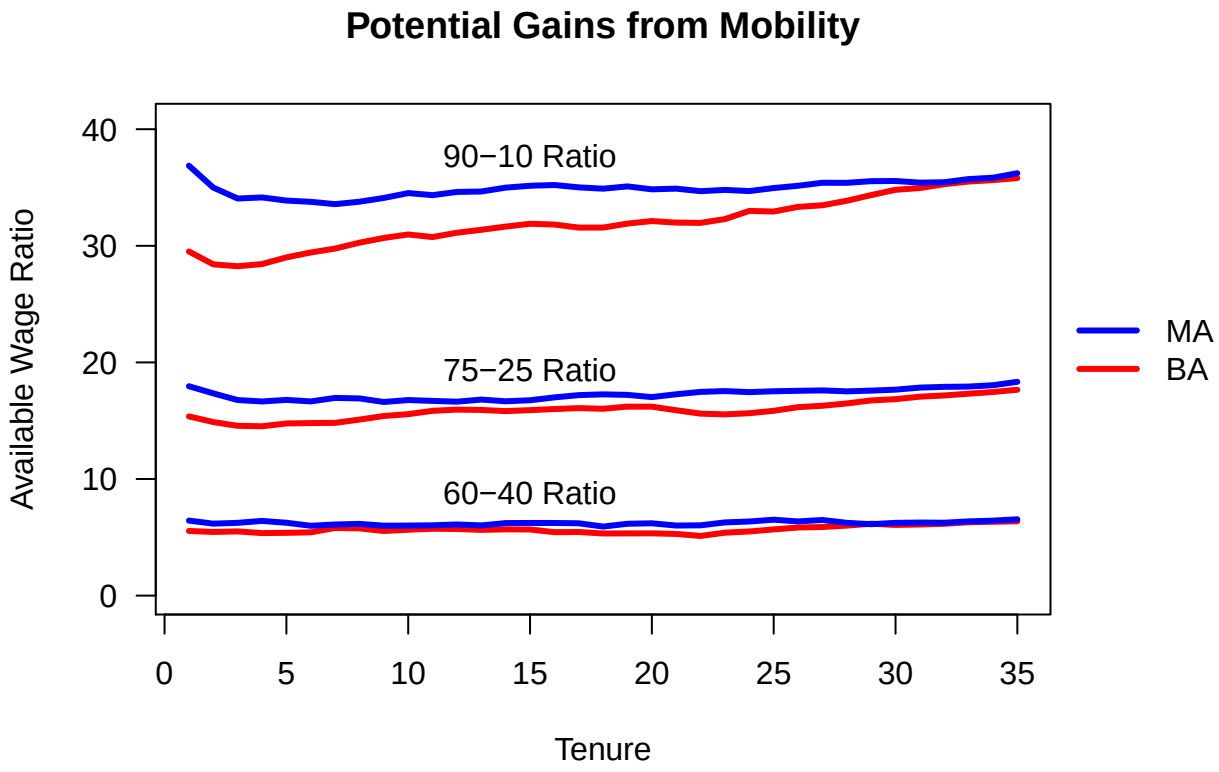

Figure 2.6: How Much Do Teachers Stand to Gain from Changing Districts throughout Their Careers?

discounting the average annual gain for Master's-certified teachers at 6% and adding over 20 years, this means roughly $100,000 is on the table; results are more dramatic for teachers at districts further in the tails of the wage distribution.

Attempting to isolate the influence of district characteristics on wage effects, HKR suggest comparing the differential leverage of residual wages (the residuals being the unexplained part of a Mincer-type regression) to get a more focused estimate of the association between wages and mobility[30]. I run a similar regression, but evaluate separate regressions not just for each level of experience, but also for each certification track. This leads to a boost in the overall fraction of explained variance from 60% cited by HKR to 87% here; as in HKR, other included covariates are consistently significant, suggesting their strong independent correlation with salary levels.

Unlike HKR, I find the demographic-independent wage differentials to be no more important than the uncontrolled raw wages, with the predicted wage improvement amounting to 0.8%. In further contrast to HKR, I find a positive relationship between experience and residual wage differentials, with mid-career district switchers experiencing roughly 1.4% higher wages upon arrival to their new employer, by contrast to the null relationship for probationary teachers. This pattern is consistent across the dimension of certification which was ignored by HKR, suggesting the opposite

[30]HKR mention they failed to adjust the standard errors associated with the adjusted wage differentials to account for the fact that they involve residuals from a regression. I explored accounting for this by bootstrapping the regression through resampling teachers and recalculating residuals, but little changes as a result, so I present the naive standard errors for simplicity.

result cannot be attributed to bias introduced by movement patterns of Bachelor's- vs. Master's-certified instructors.

Student demographic differentials are very important for predicting teacher turnover, a finding which held in Texas as it does in Wisconsin. Most distinguished in all experience classes and for both genders are changes in measures of student performance, student poverty and the percetnage of black students – district switchers end up at schools with 5% more students at grade level overall, an effect which is stronger for female teachers and for young teachers. They also end up on average with about 8% fewer students (school-wide) eligible for subsidized lunch and 7% fewer black students. While this finding would need to be bolstered with experimental or quasi-experimental evidence, it hints at the potentially limited scope of teacher labor market policies intended to ameliorate teacher supply problems in hard-to-serve districts – schools can much more easily exert influence over their compensation policies than they can dictate their student bodies, but the latter appears more efficacious (see Fulbeck 2014 and Glazerman et al. (2013)).

Long-Distance Moves

One major aspect of teacher mobility glossed over by HKR is geographic separation. A wide variety of frictions may be geospatially-related or -generated – social and professional networks tend to be concentrated locally; there are typically substantial fixed costs involved in moving (real estate closing fees, moving expenses, etc.); prefer-

	Men by Experience Class			Women by Experience Class			All Teachers	0-9 Years
	1-3 years	4-6 years	7-11 years	1-3 years	4-6 years	7-11 years		
Base year salary (log)	0.020	0.012	-0.008	-0.019	0.009	0.009	-0.002	
	(0.015)	(0.023)	(0.033)	(0.012)	(0.014)	(0.022)	(0.007)	
Adjusted salary (log)	0.016	-0.032	-0.002	-0.005	0.002	0.014	-0.000	
	(0.011)	(0.017)	(0.031)	(0.010)	(0.010)	(0.015)	(0.005)	
Percent proficient	2.1%	3.1%	1.2%	4.3%	2.9%	4.7%	3.5%	
	(1.0%)	(1.2%)	(1.5%)	(0.6%)	(1.0%)	(1.3%)	(0.4%)	
Percent Hispanic	-0.7%	0.3%	-0.3%	-1.3%	-1.2%	-1.1%	-1.0%	
	(0.5%)	(0.6%)	(0.8%)	(0.3%)	(0.4%)	(0.6%)	(0.2%)	
Percent black	-1.7%	-0.4%	-2.1%	-4.3%	-2.8%	-5.0%	-3.3%	
	(1.3%)	(1.2%)	(1.6%)	(0.9%)	(1.3%)	(1.8%)	(0.5%)	
Percent subsidized lunch	-5.9%	-5.6%	-2.6%	-7.2%	-3.8%	-6.1%	-5.9%	
	(1.6%)	(2.0%)	(3.0%)	(1.0%)	(1.5%)	(1.9%)	(0.6%)	

Table 2.4: Average Change in Salary and District Student Characteristics (and Standard Deviations) for Teachers Changing to a District More than 50 Miles Away, by Gender and Experience

ences may depend on climate/geography; and so on. As a first pass at exploring how long-distance moves may differ in nature from those over short distances, I reproduce in Table 2.4 the analysis of Table 2.3 for only those moves where the distance between the origin and destination school exceeded 50 miles (a distance deemed sufficient to likely entail a physical move rather than simply an adjusted commute).

The preeminent distinction of long-distance moves is moderation – all average demographic differentials moderate towards zero, suggesting a diminution of the importance of these aspects in this population. The noteworthy exception to this trend is among probationary teachers – young males experience wage increases in an uprooting move, while young females experience declines for long moves. More detailed data would be needed to explore the mechanism at work behind this observation (in particular, none of the differences – male vs. female or short- vs. long-distance moves – have p values below .05), but one explanation is a higher willingness among bachelors to change scenery completely, while younger women may tend to be married and moving with their partners. In any case, the overall importance of wages in long-distance moves is close to zero, suggesting wage differentials are either of secondary or tertiary concern in the associated decision processes, or that there is unsufficient heterogeneity in wages at such distances to generate enough moves so motivated, though the case of young male teachers does weaken the latter explanation.

Supply and Demand for Subject Specialists

Another source of heterogeneity about which HKR have little to say is subject specialty. While it is true that all teachers on a given contract are typically paid independently of the subject they teach, teaching a hard-to-staff subject should lead to more bargaining power in the labor market (as such teachers are less easily replaced), so I would expect such teachers to transition to more attractive positions upon moving. Fully accounting for the demand side of labor markets would bestow higher confidence in results which ultimately depend on the strategic interaction of the two sides.

I are aided in trying to explore this aspect of the teacher labor market by the public availability of annual technical reports from DPI about various aggregate indicators for the health of supply and demand for educators in Wisconsin (the last published edition is Fischer, Swanger, and Skoning 2009). In addition to providing counts for the number of educators graduating from the in-state education programs broken down by subject area, the report uses a survey distributed to district administrators to give a score (based, for example, on the market tightness – applications per vacancy) in each Cooperative Educational Service Agency (CESA, the administrative unit for districts between the school district and DPI) rating the need for educators in various subject areas, including those in my study sample, Math, Reading, and Elementary.

Both Reading and Elementary are chronically over-supplied throughout the state, whereas the demand for math teachers varies considerably. In a given year, the market tightness for the former two subjects is roughly twice that in Math (e.g., it

	1-3 years		4-6 years		7-11 years		All	
	Non-Math	Math	Non-Math	Math	Non-Math	Math	Non-Math	Math
Base year salary (log)	-0.069	0.031	0.007	0.015	0.025	-0.005	0.001	0.011
	(0.030)	(0.042)	(0.019)	(0.053)	(0.015)	(0.031)	(0.012)	(0.022)
Adjusted salary (log)	-0.008	0.068	0.031	0.047	0.034	0.019	0.025	0.041
	(0.024)	(0.024)	(0.014)	(0.047)	(0.013)	(0.017)	(0.009)	(0.015)
Percent proficient	10.8%	8.5%	6.2%	1.4%	4.4%	2.2%	6.2%	3.7%
	(2.0%)	(3.4%)	(1.3%)	(2.5%)	(1.0%)	(1.9%)	(0.8%)	(1.5%)
Percent Hispanic	-4.2%	1.4%	-2.1%	-1.1%	-0.8%	-1.4%	-1.9%	-0.6%
	(1.0%)	(1.7%)	(0.6%)	(1.5%)	(0.5%)	(1.3%)	(0.4%)	(0.9%)
Percent black	-15.0%	-11.8%	-6.6%	0.4%	-5.1%	-4.2%	-7.5%	-4.9%
	(3.2%)	(5.3%)	(1.8%)	(3.3%)	(1.2%)	(2.5%)	(1.1%)	(2.1%)
Percent subsidized lunch	-17.6%	-7.3%	-7.9%	-2.9%	-4.9%	-4.8%	-8.4%	-4.9%
	(3.3%)	(5.7%)	(2.3%)	(5.2%)	(1.6%)	(3.4%)	(1.3%)	(2.6%)

Table 2.5: Average Change in Salary and District Student Characteristics (and Standard Deviations) for Teachers with Master's Degrees Changing Districts, by Subject Area and Experience

was 67.43 for Elementary, 28.65 for English/Speech/Theater/Journalism, and 24.22 for Mathematics). As a result, I expect to see some heterogeneity in labor market success of specialists in Math as compared to the other teachers in my sample. Table 2.5 explores some of the basic insights on subject matter heterogeneity. To mitigate the potential for degree holdings to skew results, I focus on Master's holders and obfuscate gender differences for brevity. Actually, there is little in this table to support the hypothesis that math teachers are given a substantial advantage in the labor market – math teachers earn more (both in nominal and beyond-demographic pay), but this result is not significant. Further, English teachers are advantaged in ending up at less economically disadvantaged and higher-performing districts.

Table 2.6, which parallels Table 4 of HKR, again uncovers a labor market functioning similar to that in Texas. In particular, while HKR find Large Urban - Suburban district switchers penalize themselves in pay but are rewarded in demographic-adjusted pay, Wisconsin teachers lose out on both measures when leaving Large Urban districts, albeit the residual pay penalty is much lower than that of nominal pay. This difference does not appear to be attributable to HKR's exclusion of certification as a conditioning variable, as the pattern here differs insignificantly by degree.

The other results of HKR are confirmed in even more dramatic fashion. There is strong evidence of selection on the student performance metric, which does vary quite widely in suburban districts. Teachers leaving Milwaukee tend to end up at districts with 38% more students deemed to be at grade level on the state standardized test.

	District Average Characteristics		Campus Average Characteristics	
	Large Urban to Suburban	Suburban to Suburban	Large Urban to Suburban	Suburban to Suburban
Base year salary (log)	-0.056	0.018	—	—
	(0.012)	(0.007)		
Adjusted salary (log)	-0.004	0.011	—	—
	(0.005)	(0.006)		
Average Student Characteristics				
Percent proficient	37.9%	0.9%	35.1%	0.1%
	(0.6%)	(0.4%)	(1.2%)	(0.6%)
Percent Hispanic	-11.3%	-0.6%	-7.3%	-0.4%
	(0.4%)	(0.2%)	(1.3%)	(0.2%)
Percent black	-56.9%	-0.6%	-59.7%	-0.5%
	(0.8%)	(0.3%)	(1.8%)	(0.4%)
Percent subsidized lunch	-55.7%	-1.7%	-61.1%	-1.6%
	(1.2%)	(0.5%)	(1.3%)	(0.7%)

Table 2.6: Average Change in Salary and in District and Campus Student Characteristics (and Standard Deviations) for Teachers with 1-10 Years of Experience Who Change Districts, by Community Type of Origin and Destination District

On the other hand, teachers leaving Large Urban districts (i.e, Milwaukee) for the suburbs experience a precipitous drop of 57% black students and 56% subsidized lunch eligibility. This is practically a tautological result, as the student demographics outside of urban areas in Wisconsin are pretty uniformly non-minority – about 90% of suburban districts have fewer than 10% black students, and about 60% have fewer than 2% black students, whereas Milwaukee is about 60% black. Similarly, teachers leaving Milwaukee for the suburbs have little choice but to end up in a district with far fewer economically disadvantaged students – whereas 73% of Milwaukee students are eligible, the median percentage in suburban schools is 12%.

The direction of these effects are preserved among suburban-to-suburban moves, suggesting the importance of these factors even in areas where there is a wider array demographically of destination districts. I also find evidence of selection into economically better-off districts among suburban switchers, but the magnitude of this difference is attenuated with respect to that reported by HKR. I do not find patterns of selection on student performance as strongly as was found in HKR. This may be a reflection of the crudeness of the proficiency measure as compared to the more variable raw scale score measures used by HKR. Lastly, I confirm the finding of HKR that there does not appear to be evidence that teachers are able to select into the more desirable schools within their target districts – the differences in campus-level characteristics are almost identical to the differences in district-level characteristics. This is likely a reflection of supply-side constraints, as the choicest appointments in

a district may be awarded to long-serving serving teachers (promotion from within), as well as suburban districts perhaps having only a small number of schools at which to teach a given grade level/subject.

	Between District Moves		Within District Moves	
	Black Teachers	Hispanic Teachers	Black Teachers	Hispanic Teachers
Percent proficient	10.7%	8.0%	2.7%	2.2%
	(3.4%)	(5.6%)	(0.9%)	(1.3%)
Percent Hispanic	3.2%	-14.8%	1.0%	-7.7%
	(1.4%)	(7.3%)	(0.9%)	(2.3%)
Percent black	-21.1%	-0.6%	-2.1%	-0.3%
	(5.0%)	(5.0%)	(1.4%)	(2.0%)
Percent subsidized lunch	-19.1%	-15.5%	-3.5%	-4.7%
	(7.7%)	(6.7%)	(0.8%)	(1.3%)
Number of teachers	81	37	638	228

Table 2.7: Average Change in District and Campus Student Characteristics (and Standard Deviations) for Black and Hispanic Teachers with 1-10 Years of Experience who Change Campuses

HKR examine the state of Texas, which features substantially more ethnic heterogeneity than does Wisconsin. As a result, they are better-equipped to identify

heterogeneity in preferences by teacher ethnicity. In Wisconsin, however, only 2,372 of the 49,325 teachers are non-white, so my results are underpowered relative to HKR. For completeness, Table 2.7 presents these results, which parallel HKR Table 5. Given how few observations I have of black or Hispanic teachers switching districts, I eschew any temptation to interpret these results. Only black switchers within districts provide enough records to interpret meaningfully. In Wisconsin, I find that, in contrast to white within-district switchers, black teachers tend to migrate to economically better-off and higher-performing schools (white teachers also select on percentage of black students). This could simply be a reflection of differences in initial district choice by black vis--vis white teachers – the median proficiency at a black teacher's first district is 36%, compared to 64% for white teachers (71% and 22% for reduced lunch eligibility, respectively).

To the end of examining heterogeneity in the impact of school and district characteristic differentials on teacher mobility, HKR present their Table 6, which breaks down the three exit rates for each (weighted) quartile of the covariate distribution. I replicate that analysis here in Table 2.8. Saliently, my results for the correlation of school characteristics for within-district movers are qualitatively identical to those found in Texas and similar in magnitude, which gives a stronger indication that I have identified some fundamental nonpecuniary mechanisms driving sorting among schools in a district.

Differences with respect to the results in Texas begin to emerge for the other

Quartile of Distribution	Probability Teachers Move to New School within District	Probability Teachers Move to New District	Probability Teachers Exit Public Schools
Residual salary			
Highest	—	1.5%	4.1%
3rd	—	1.8%	5.0%
2nd	—	1.8%	4.9%
Lowest	—	1.9%	4.1%
Percent proficient			
Highest	4.5%	1.9%	4.2%
3rd	4.6%	2.3%	4.2%
2nd	5.2%	1.7%	4.4%
Lowest	6.1%	2.1%	4.6%
Percent eligible for reduced-price lunch			
Highest	7.1%	2.1%	5.3%
3rd	5.6%	1.7%	3.8%
2nd	4.1%	2.0%	3.9%
Lowest	3.6%	2.2%	4.4%
Percent Black			
Highest	7.3%	2.1%	5.9%
3rd	4.9%	1.6%	4.2%
2nd	4.7%	1.9%	3.8%
Lowest	3.4%	2.4%	3.4%
Percent Hispanic			
Highest	7.6%	1.8%	5.5%
3rd	4.4%	2.0%	4.1%
2nd	4.3%	2.0%	4.0%
Lowest	4.0%	2.3%	3.7%

Table 2.8: School Average Transition Rates by Distribution of Residual Teacher Salary and Student Demographic Characteristics (data weighted by number of teachers in school)

destinations of school leavers (other districts and other professions). As noted in Table 2.1, overall rates of switching districts are quite low compared to Texas and national averages; conditional on this, the patterns of movement by quartile of residual salary exhibit a similar pattern to that in Texas, with teachers in the lowest quartile about 28% more likely to change districts than teachers in the highest residual pay quartile. By contrast to HKR, however, who found the opposite association, I find the same trend (at attenuated magnitudes) with respect to leaving Wisconsin public schools, suggesting salary considerations are also important for teachers considering options outside of public school teaching (or in other states).

I also find fairly strong patterns in quitting associated with subsidized lunch eligibility and with the ethnic makeup of schools, with teachers at the most economically advantaged schools 8% less likely to exit teaching; similar numbers obtain for both the quantity of black and of Hispanic students. For teachers moving within districts, I observe similar patterns.

Regression Results

Having identified some key patterns in moments of the data, I now move on to try and separate the confounding effects of each of these and other factors in affecting teacher turnover with the aim of identifying more fundamentally the association between salient district and school characteristics on teacher turnover. Table 2.9 provides the main coefficients of interest from a simple linear probability regression model

	Teacher Experience				
	1-3 years	4-6 years	7-11 years	12-30 years	>30 years
First year base salary (log)	0.03	-0.09^{**}	-0.04^{*}	0.01	-0.12
	(0.03)	(0.03)	(0.02)	(0.01)	(0.06)
First year base salary (log) * female	-0.07^{*}	0.09^{**}	0.02	-0.02	0.12^{*}
	(0.03)	(0.03)	(0.02)	(0.01)	(0.06)
Campus average student characteristics					
Percent proficient	-0.10^{*}	0.03	-0.02	0.00	-0.07
	(0.05)	(0.04)	(0.02)	(0.01)	(0.06)
Percent eligible for subsidized lunch	-0.09^{**}	-0.07^{**}	-0.07^{***}	-0.01	0.07
	(0.03)	(0.03)	(0.02)	(0.01)	(0.04)
Percent Black	0.04	0.22^{***}	0.15^{***}	0.09^{***}	0.14^{*}
	(0.04)	(0.04)	(0.03)	(0.02)	(0.07)
Percent Hispanic	0.13^{*}	0.16^{***}	0.04	-0.04^{*}	-0.19^{*}
	(0.06)	(0.05)	(0.03)	(0.02)	(0.08)
Interactions					
Black * percent Black	-0.21^{**}	-0.13	-0.03	-0.03	-0.21
	(0.08)	(0.07)	(0.05)	(0.04)	(0.12)
Hispanic * percent Black	-0.19^{***}	-0.19^{***}	-0.12^{*}	-0.12^{**}	0.02
	(0.06)	(0.05)	(0.05)	(0.04)	(0.34)
Black * percent Hispanic	0.14	-0.23	-0.14	0.03	-0.53
	(0.25)	(0.23)	(0.14)	(0.11)	(0.45)
Hispanic * percent Hispanic	0.14	-0.07	-0.21	0.26	0.61
	(0.28)	(0.23)	(0.20)	(0.19)	(1.07)
Observations	33,108	30,244	43,509	98,753	15,217

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2.9: Estimated Effects of Starting Teacher Salary and Student Demographic Characteristics on the Probability that Teachers Leave School Districts, by Experience (linear probability models; Huber-White standard errors in parentheses)

predicting leaving a district (i.e., either switching districts or exiting teaching); this corresponds to HKR Table 7.

By contrast to the strength implied in earlier results, the importance of student achievement has dwindled in the regression specification, and only comes out as independently significant for probationary teachers. The same goes for base salary differentials – in contrast to HKR, the evidence I find in favor of an independent influence of salary on turnover rates is sparse and concentrated among young teachers[31]. This does not appear to be due to imprecision – the magnitude of HKR's standard errors follows closely those found for the Wisconsin data, despite my smaller sample sizes.

HKR also found little independent evidence in favor of student economic status factoring in to teachers' mobility decisions, but I find fairly consistent support for the importance of subsidized lunch eligibility prevalence. As mentioned above, it is possible that the crude nature of the proficiency measure is only weakly identified, and that some of the unaccounted for part of student performance is being captured in other coefficients, especially subsidized lunch eligibility and student race. Even more compelling would be to associate student performance (and other school/district-level characteristics) more finely with the set of students actually faced by a given teacher.

[31]HKR also mention results not printed in their paper suggesting a paucity of evidence suggesting class size is an important factor in teacher turnover decisions; I give tepid support to this statement, as class size does indeed appear to be related to turnover, but somewhat weakly and only for younger teachers.

The results in HKR about the differential effects of student body makeup are largely similar to those I find in Wisconsin. White and nonwhite teachers have opposite and significant correlations between the quantity of minority students in their origin district and their likelihood of leaving it. These differential results tend to modulate towards zero with experience, regardless of teacher or student race category, and suggest a degree of assortative matching on ethnicity among districts in Wisconsin (though the patterns for whites differ sharply from those of nonwhites, the patterns for black and Hispanic teachers are hard to distinguish).

To account in a rudimentary way for district-specific hiring policies, HKR move on to their Table 8 which repeats Table 7 (my Table 2.9) with district fixed effects. HKR note that the patterns in responsiveness to wages are the same, though attenuated; that coefficients involving student ethnicity are qualitatively unaffected; and that schools with high achievement continue to exhibit lower propensities for turnover. My results, presented in Table 2.10, are similar in that they closely resemble the results without fixed effects, but with noted attenuation and weaker precision.

The most notable difference relative to Table 2.9 is the general weakening of results regarding the importance of student characteristics for white teachers. While partially attributable to a decline in precision, this adjustment suggests much of the discovered correlation between student characteristics and exit probability for white teachers can be chalked up to district-to-district heterogeneity in preferences or hiring policies.

	Teacher Experience				
	1-3 years	4-6 years	7-11 years	12-30 years	>30 years
First year base salary (log)	0.01	-0.13^{***}	-0.05^{**}	0.00	-0.17^{**}
	(0.03)	(0.04)	(0.02)	(0.01)	(0.07)
First year base salary (log) * female	-0.07^{*}	0.09^{**}	0.03	-0.02	0.13^{*}
	(0.03)	(0.03)	(0.02)	(0.01)	(0.06)
Campus average student characteristics					
Percent proficient	-0.12	-0.04	-0.10^{*}	0.01	-0.09
	(0.08)	(0.06)	(0.04)	(0.02)	(0.10)
Percent eligible for subsidized lunch	-0.07	-0.05	-0.03	-0.02	-0.21^{*}
	(0.08)	(0.06)	(0.04)	(0.02)	(0.11)
Percent Black	0.29	0.46	0.54^{**}	0.03	0.38
	(0.28)	(0.26)	(0.18)	(0.10)	(0.48)
Percent Hispanic	0.04	0.04	-0.23^{*}	-0.07	0.13
	(0.18)	(0.15)	(0.09)	(0.05)	(0.25)
Interactions					
Black * percent Black	-0.19^{*}	-0.15^{*}	-0.06	-0.04	-0.17
	(0.08)	(0.07)	(0.05)	(0.04)	(0.12)
Hispanic * percent Black	-0.17^{**}	-0.18^{***}	-0.12^{*}	-0.11^{**}	-0.09
	(0.06)	(0.05)	(0.05)	(0.04)	(0.33)
Black * percent Hispanic	0.06	-0.15	0.02	0.09	-0.73
	(0.26)	(0.25)	(0.15)	(0.11)	(0.48)
Hispanic * percent Hispanic	0.14	-0.04	-0.12	0.26	0.26
	(0.28)	(0.23)	(0.21)	(0.20)	(1.10)
Observations	33,108	30,244	43,509	98,753	15,217

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2.10: Estimated Effects of Starting Teacher Salary and Student Demographic Characteristics on the Probability that Teachers Leave School Districts with District Fixed Effects, by Experience (linear probability models; Huber-White standard errors in parentheses)

Local Wage Ratio

Perhaps a better measure of the influence of a teacher's wage on their propensity to move is their potential gains from changing to nearby districts. Table 2.11 presents results of a specification paralleling that in Table 2.10, save for the replacement of the initial Bachelor's salary as a predictor with a measure of a teacher's local relative wage. Specifically, the local relative wage is defined as the ratio of a teacher's next scheduled wage to the wage ceiling at districts within 50 miles of their current district (excluding their own) The wage ceiling is calculated as the maximum wage at the teacher's subsequent level of experience and certification in such districts. If this measure exceeds one, a teacher is getting paid more for their current qualifications than is possible locally; otherwise, they stand to gain in pay from switching to at least one school locally. This relative local wage measure is an even poorer predictor of teacher churn than are local wage levels. Table 2.11 bolsters evidence that student characteristics are more important than wage differentials for teachers considering changing districts[32].

Finally, the conflation of switching districts and exiting teaching may mask important heterogeneity between these two choices. To separate these competing exit risks,

[32]Some of the other coefficients in Table 2.11 have changed somewhat substantially relative to those found in Table 2.10. This is largely attributable to a different subpopulation of teachers included in the results, as evidenced by the change in sample size between the two tables. This reflects differential missingness of the wage measures. The wage coefficients are qualitatively robust to selecting a fixed population to estimate the models for the two tables.

	Teacher Experience				
	1-3 years	4-6 years	7-11 years	12-30 years	>30 years
Local Relative Wage	0.00	0.01	−0.02	−0.02**	−0.01
	(0.03)	(0.02)	(0.02)	(0.01)	(0.01)
Local Relative Wage * female	0.00	0.00	0.03	0.01**	0.01
	(0.03)	(0.02)	(0.02)	(0.00)	(0.01)
Campus average student characteristics					
Percent proficient	−0.03	0.04	−0.01	−0.01	−0.01
	(0.06)	(0.04)	(0.02)	(0.01)	(0.00)
Percent eligible for subsidized lunch	−0.12*	−0.00	−0.02	0.00	−0.00
	(0.06)	(0.04)	(0.02)	(0.01)	(0.01)
Percent Black	0.18	0.31*	0.21	0.08	0.00
	(0.23)	(0.15)	(0.13)	(0.05)	(0.01)
Percent Hispanic	0.09	0.09	−0.06	−0.02	0.00
	(0.13)	(0.10)	(0.06)	(0.02)	(0.02)
Interactions					
Black * percent Black	−0.11**	−0.05*	−0.04***	0.00	0.00
	(0.04)	(0.02)	(0.01)	(0.01)	(0.00)
Hispanic * percent Black	−0.06*	−0.06*	−0.08*	−0.00	0.00
	(0.03)	(0.02)	(0.04)	(0.00)	(0.00)
Black * percent Hispanic	0.25	0.25**	0.12	0.00	−0.00
	(0.14)	(0.08)	(0.06)	(0.04)	(0.00)
Hispanic * percent Hispanic	0.12	0.06	0.15	0.02**	0.00
	(0.11)	(0.07)	(0.12)	(0.01)	(0.01)
Observations	27,045	25,285	36,364	83,825	10,084

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

Table 2.11: Estimated Effects of Relative Local Wage and Student Demographic Characteristics on the Probability that Teachers Leave School Districts with District Fixed Effects, by Experience (linear probability models; Huber-White standard errors in parentheses)

	Teacher Experience			
	1-3 years	4-6 years	7-11 years	12-30 years
I. Switch Districts				
First year base salary (log)	0.66	−0.27	−0.81	1.36
	(0.53)	(0.74)	(0.71)	(0.91)
First year base salary (log) * female	−1.12*	−0.10	−0.36	−1.83*
	(0.52)	(0.73)	(0.73)	(0.91)
Percent proficient	−1.38*	0.16	−1.05	1.20
	(0.60)	(0.80)	(0.89)	(1.03)
Percent eligible for subsidized lunch	−0.57	−0.91	−1.98**	−0.62
	(0.40)	(0.57)	(0.64)	(0.74)
Percent Nonwhite	0.15	2.11***	2.87***	3.19***
	(0.40)	(0.54)	(0.58)	(0.66)
Nonwhite * percent Nonwhite	−2.45***	−3.68***	−3.08**	0.11
	(0.56)	(0.87)	(1.04)	(1.29)
II. Exit Teaching				
First year base salary (log)	0.27	−2.29***	−0.76	0.05
	(0.52)	(0.57)	(0.64)	(0.44)
First year base salary (log) * female	−0.63	2.43***	0.73	−0.32
	(0.51)	(0.61)	(0.65)	(0.44)
Percent proficient	−0.46	0.42	0.02	0.12
	(0.57)	(0.68)	(0.71)	(0.45)
Percent eligible for subsidized lunch	−0.88*	−1.50**	−2.31***	−0.57
	(0.39)	(0.49)	(0.53)	(0.31)
Percent Nonwhite	1.24***	3.17***	2.36***	0.90**
	(0.35)	(0.42)	(0.46)	(0.29)
Nonwhite * percent Nonwhite	−0.97*	−1.75***	−1.37**	−0.93*
	(0.42)	(0.45)	(0.47)	(0.38)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2.12: Multinomial Logit Estimated Effects of Teacher Salary and Student Demographic Characteristics on the Probabilities That Teachers Switch School Districts or Exit Teaching Relative to Remaining in Same District

HKR construct Table 9, which gives coefficients from a multinomial logit model with three choices – remain in district, switch districts, and exit teaching. I repeat that analysis here in Table 2.12, with the caveat that, given the sparsity in racial variation present among Wisconsin teachers, I are unable to identify the full model specified by HKR and mirrored above in Tables 2.9 and 2.10. In light of this, and in light of the apparent similarity in Wisconsin in the behavior of black and Hispanic teachers described above, I specify the multinomial logit model in terms of a more parsimonious coefficient set. Namely, I distinguish between white and nonwhite teachers and white and nonwhite students (instead of among white, black, and Hispanic students and teachers).

I continue to see little evidence favoring the salience of wage considerations for Wisconsin teachers; the strongest suggestions found here point to the importance of wages for older male teachers in exiting teaching, a result which is generally opposed to that found by HKR in Texas, where salaries were generally important, but only for the propensity to change districts. Also as in the regression specifications, the prominence of student proficiency found by HKR fails to make a notable appearance in Wisconsin.

With respect to the importance of student demographics, my results again point to the same effects found in Texas. White teachers seem to be spurred to change districts or exit teaching by highly black student populations; the reverse is true of nonwhite teachers, who can be drawn to remain in high-minority districts. Subsidized

lunch eligibility's strong effect observed in the combined specification is found here to be concentrated more among those leaving teaching than those changing districts.

Chapter 3

Active Learning Classrooms for College Calculus Instruction

Literature Review

I have identified no studies that have evaluated the effectiveness of sufficiently similar types of active learning classrooms as that studied here relative to traditional mathematics instructional methods. There is, however, a body of prior literature comparing the outcomes of online and blended instruction within both K-12 and postsecondary education.

A previous study (Deslauriers, Schelew, and Wieman 2011) that measures instructional approaches that most closely parallel those studied here reported promising findings for blended-delivery physics courses. In a semester-long randomized experiment of calculus-based physics courses at the University of British Columbia, students

who attended a course that was taught using an "interactive instructional approach based on research on learning" (Deslauriers, Schelew, and Wieman 2011, 862) were more likely to attend and be engaged in class than students attending a traditional lecture-based section. Furthermore, students in the treatment section scored 33 points higher on a post-experiment exam than students in the control group.

McGivney-Burelle and Xue (2013) perform the closest study in examining the effectiveness of flipping a single unit of an intermediate calculus course analogous in the calculus progrssion to that under study here. The study is non-random, but has the unique aspect of mitigating instructor fixed effects by having the same instructor cover the flipped unit of the course. They report high student satisfaction and engagement with this unit. Love et al. (2014) do a non-random evaluation of flipped classrooms for college linear algebra and similarly find high student satisfaction (though no significant differences in outcomes emerged). Given the non-random nature of the two studies, suspicions of selection effects and external validity abound.

Other research provides a more mixed picture. For example, in a meta-analysis of 45 studies of online and blended learning across various postsecondary settings, Means et al. (2009) found evidence of modest benefits of on-line versus traditional face-to-face models of instruction on math skills. However, the largest skills gains were found for courses that blended the two instructional formats. The authors note that several factors in addition to the mode of delivery may contribute to these differences in skills gains. In response to the Means et al. (2009) study, Jaggars and Bailey (2010) note

that the findings do not hold up for semester-length courses delivered fully online. Moreover, they note that the majority of the evidence base is limited to students who were academically prepared to complete the coursework and, thus, have limited applicability to those most challenged in traditional math classes. Andrews et al. (2011) find instructor experience to be a key factor in the success of actively-taught classrooms in their examination of college biology courses selected from universities nationwide at random.

Two recent studies found mixed evidence of the effectiveness of online versus traditional math instructional strategies. Holding several student characteristics constant, Xu and Jaggars (2011) and Xu and Jaggars (2013) reported evidence suggesting that failure and withdrawal rates were higher and persistence rates were lower for students in community college online courses than for students in courses taught face-to-face, while Xu and Jaggars (2011) reported no evidence of differences in outcomes between students in on-line versus blended instruction courses. Moreover, Xu and Jaggars (2013) reported evidence suggesting that on-line instruction may be especially disadvantageous for males, Black students, and students with lower levels of academic preparation.

Yet, there is evidence that context matters. For example, a recent randomized control trial of rural middle school students taking Algebra I online found evidence that these students outperformed their counterparts who did not have access to the on-line course on their state mathematics assessments, quite likely due to higher level

of access to any Algebra instruction, rather than as a consequence of the particular instructional mode (Heppen et al. 2011). Berlinski and Busso (2015) implement a randomized-controlled trial in Costa Rican secondary schools to foster active student engagement with math and find significantly negative outcomes for treated students, an outcome which they attribute to a deterioration in the quality of student-teacher interactions in this group.

Study Setting and Instructional Contrast

The study focuses on strategies for teaching Introductory Calculus to college students – a course that covers integration by parts, basic differential equations, and sequences and series through Taylor Series representation of functions. This is a required course for most college freshmen and a gateway course for prospective STEM students as it is a prerequisite for most intermediate-to-advanced courses in engineering, math, economics, chemistry and physics.

The setting for the study is a mid-sized private university where about 20 percent of its freshmen enroll in the course each fall. The total study sample includes all freshmen students who enrolled in the course in the fall terms of the 2014-15 or the 2015-16 academic years (N = 1,490). The course is taught each spring, as well, but Spring enrollments are very modest relative to Fall enrollments and the Spring classes tend to be populated by students who needed a semester of preparatory mathematics instruction and students who failed the course previously. As such, I exclude them

from comparison against students in Fall sections, though understanding the impact of the intervention separately among Spring students is a topic of future interest.

The study examines the implications for student course performance of two contrasting instructional strategies for the course. All sections of the calculus course under study had identical course goals, used a common textbook, and administered the same mid-term and final exams. All instructors were expected to spend a total of 42 hours in the lecture sessions of the course (i.e., either three 1-hour or two 1.5 hour lectures and a one-hour recitation per week), regardless of the instructional format. The focal differences in the instructional strategies pertained to the classroom setting, what work students are expected to do outside of the classroom, and how students and instructors use class time.

Traditional Instructional Approach

The "traditional" mode of calculus instruction centers on the lecturer, who prepares a talk on a topic each class period and presents a new tool or concept to the students with slides, derivations on the chalkboard, and verbal feedback. Students are typically seated in tiered seating environments, and the audience may be between 60 and 100 students. Outside of the classroom, students are assigned problem sets (usually on a weekly basis) that typically include a selection of exercises from the course's required textbook. Students are free to form study groups of their own accord; attendance of the lectures may or may not be mandatory, depending on the instructor's

preferences[33].

The Active Learning Classroom

The alternative instructional strategy being examined in this study is what I refer to as an "active learning" classroom. In this setting, a significant proportion of class time is used for instructor-moderated group work exploring and/or solving problems. To facilitate group work, students in the active classrooms sit at mid-sized tables (sitting 4-8 students per table). Generally, the class begins with an overview of the scheduled topic, under the assumption that students will have read the assigned text and viewed assigned online materials. A substantial portion of class time is used for instructor-assigned problem sheets covering exercises intended to practice the concepts and mechanics of a new tool or topic. Students work with other students to complete the problem sheet in class, as the instructor and teaching assistants roam the room to monitor and address questions. In cases where the instructor or teaching assistants note student challenges with conceptual issues, he/she may interrupt the group work to provide some general guidance – e.g., delineating the details of the snag and offering guidance in how to think about the issue.

[33]This course also features weekly recitation sessions with a teaching assistant; classroom observations revealed these exercise-oriented hour-long reviews to be sufficiently similar across the pedagogical modes that I eschew focusing on them in this analysis.

Theory of Change

The core motivation for encouraging an active learning strategy is that students will gain mathematical fluency more readily if they are engaged more tangibly with the material. This may be especially true for introductory-level courses where becoming comfortable with manipulating basic symbols and getting an intuitive understanding of concepts through constant exposure to canonical examples following familiar patterns can be seen as a fundamental step to success in higher-level courses that require more and more abstract thinking. Such forced repeated exposure to worked examples in a setting alongside peers is precisely the focus of the active learning instructional mode under study.

Of course, group completion of assignments has long been an option for aspiring students of calculus. The group setting designed for active learning sections differs in several meaningful ways from the *ad hoc* association of students typical of more traditional learning sections. First, students are aided by having on-hand the subject experts (the course instructor and their teaching assistants). Observing in real time how students struggle with the mechanics of implementing new concepts allows the professor to address misunderstandings sooner than is possible in a setting where students do not apply new tools on their own for perhaps several days after a lecture.

Active learning classroom groups can also be designed with more agency – with some experience, an instructor may come to know which types of groups tend to succeed and which combinations are more prone to distraction (e.g., the professor may

come to have a preference for "tracking" students within the class – assigning students of like ability level to the same group – or to encourage mixing – perhaps having one student with a higher level of mastery than their groupmates serving as local luminary)[34]. Devolving the group formation duty to the students themselves introduces incentives which may run counter to maximizing conceptual understanding.

Lastly, related to the competing interests of *laissez-faire* groups, by establishing a regular and formal exposure to this setting, students are likely to engage more openly with their own mathematical erudition and become more comfortable describing their math verbally to others, a key communication skill which may be otherwise lost or longer in developing, particularly for students who are not naturally inclined to form groups in such a course (whether out of confidence or timidity). Anecdotally, this type of communication – explaining a new concept that is just beyond the reach of a similarly situated peer – is a powerful way of moving from understanding to mastery of new topics.

It is precisely this numeracy and comfort with communicating mathematical ideas (and mathematical struggles) that is the intended mechanism for long-term advancement in STEM generated by active-learning instruction. Forcing students to spend three hours per week actively partcipating in discussions of math topics can lead to quantitatively confidence, which is in turn a veritable prerequisite for STEM success.

[34]In informal talks with active learning instructors, the group assignment mechanism was to date used somewhat sparingly, probably owing to the recency with which the instructors began utilizing this mode.

Study Design and Data

The implications for student performance of the instructional strategy used is being tested in two ways – using a randomized controlled trial for the 2015-16 cohort and using quasi-experimental methods for the 2014-15 and 2016-17 cohorts. This chapter focuses only on the student sample from the 2015-16 randomized controlled trial. For the randomized controlled trial, assignment was logistically constrained to be blocked at the timeslot level to allow the course to fit in predictably with students' multifarious scheduling constraints as it would have absent my intervention. To that end, I coordinated with the course registration office to create 'holding sections' to which students could register; what was visible to students during the enrollment period is depicted for a representative lecture time slot by part A of Figure 3.1.

The students registered for each of the three holding sections were then randomly assigned to a lecturer, two of which were available for each holding section, one "traditional" and one "active" (section C in Figure 3.1), for a total of six lecture sections for the class as a whole. These holding sections had the correct time of day (for both the lecture and the recitation sections, as depicted in part D of Figure 3.1), but did not reveal the format of the class or the instructor until after registration had completed[35].

[35]I could not incorporate upper-classmen (i.e., non-freshmen) to the randomization procedure. Registration was opened to them much earlier, *circa* April, whereas incoming students by that time had not necessarily even chosen their post-secondary institution. While taking this course in the Spring semester is not uncommon, doing so after first year is quite rare; as such, I simply exclude

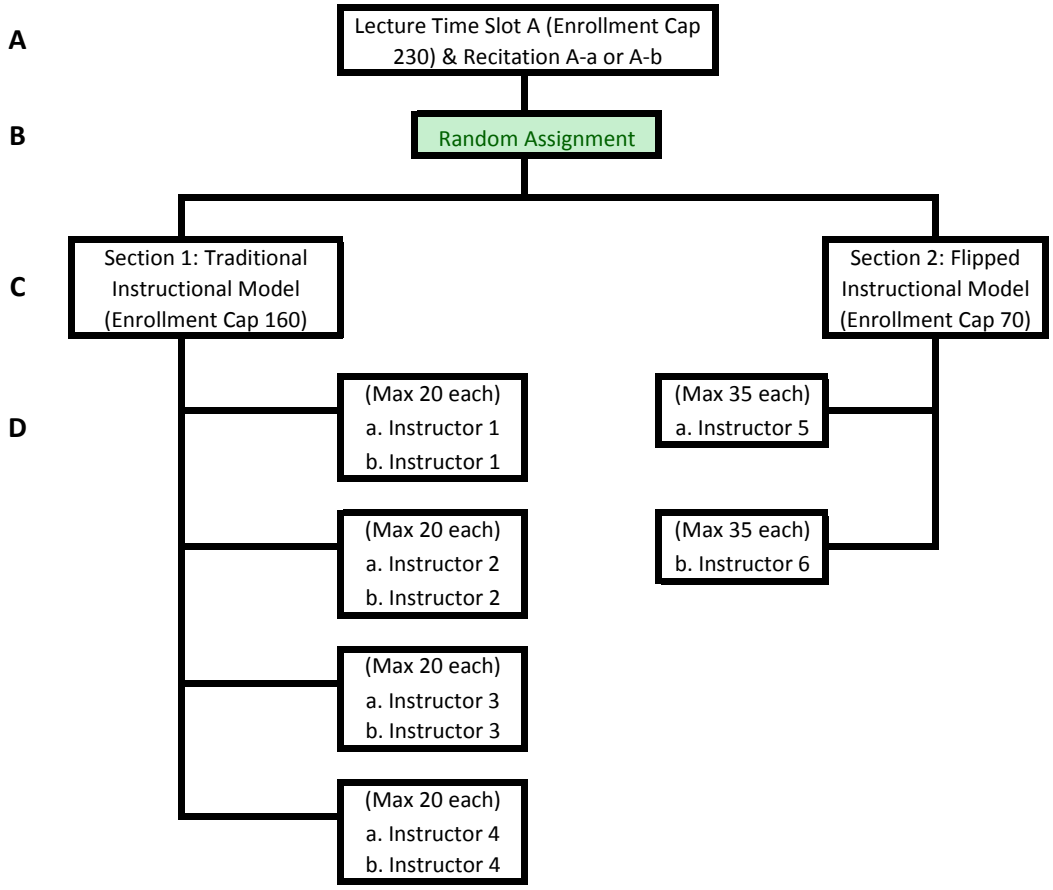

Figure 3.1: Framework for Assignment of Students to Traditional or Active Learning Sections

Time Slot	# Registered	Lecture Style	# Assigned
A	170	Traditional	120
		Active	50
B	190	Traditional	130
		Active	60
C	130	Traditional	90
		Active	40

Table 3.1: Initial Assignment to Lectures

I worked with the University Provost office, the Department of Mathematics, and other internal partners to facilitate data acquisition and exchange. I use five primary data sources – an initial pull from administrative data of registered students that I used to randomize initial enrollees to the course over the summer; an administrative cross-section of student enrollment records following the end of the add/drop deadline which delineated for each student the date they added their current lecture and recitation sections, used to identify noncompliance; a follow-up cross section which provides demographic data on course completers, as well as their final course grade[36]; formal observations scheduled by my study team of active learning and traditional

these students from my analysis.

[36]I should note that because I only have two cross-sectional pulldowns of the data, there is likely to be some measurement error in identifying the exact nature of attrition in the course. For example, there could be students who switched more than once, or who switched before dropping, or who added the course in one lecture mode and switched to the other.

lectures and recitation sections, including the completed protocols returned by observers; the results of diagnostic exam scores administered by the department at the beginning of the Fall term to help direct students to the appropriate math course; and raw final scores, provided by the professors themselves, which give the most accurate/comparable reflection of student performance on the common final exam[37].

I supplement student demographic data in a variety of ways to make it more digestible. First, I use the set of 200 students for whom both SAT and ACT performance are available to predict a student's missing SAT score from their ACT score (and vice versa)[38]. I also spent some effort converting the students' city of permanent residence into a broader regional indicator, including identifying global cities to try and place foreign nationals into cognate geographic groups.

Revelation of random assignment to the students took place a few weeks before the start of classes in the fall semester of 2015, at which point students were free (through the cessation of the standard university-wide add/drop period) to switch

[37]I were unable to obtain final exam scores for a small number of students that took a makeup exam and still passed the course.

[38]A cross-validated/machine-learned approach would be preferable were this a more crucial part of the analysis; for the purposes here, it was judged from a scatterplot of the scores that two simple regressions would suffice. These had R^2 0.41 (SAT from ACT) and 0.49 (ACT from SAT). More broad-based conversions were considered (see, e.g., Dorans 1999), but the set of students at the same school is likely a better comparison group to extrapolate to than is the whole country, i.e., I expect out-of-sample predictive ability to be better using the regression fit with only my study sample than it would be using a more general conversion between ACT and SAT scores.

lectures/recitations as they saw fit (as they would with any normal class for which they had registered). Numbers in Table 3.1 are illustrative of the registration and assignment process (numbers have been rounded for anonymity). The active learning classrooms are more space-intensive than traditional lecture halls, so I were logistically constrained to randomize roughly $\frac{2}{3}$ of students to the traditional sections in each time slot.

Intent to treat estimates of program impacts (detailed below) will be based on comparisons of students assigned to the two conditions. However, the interpretation of the study findings is complicated by two factors: (1) some students dropped the class after assignment; and (2) many students changed section after the start of the semester, leading to high rates of noncompliance with the randomly assigned treatment condition. In an effort to track student mobility, I track students' assigned lecture/recitation pair and pull two cross-sections of this assignment during the semester (including final assignment).

Noncompliance

As seen in Table 3.2, attrition from initial assignment was high and dropping the course was the most common outcome for this subsample – overall, 28% of initially enrolled students did not complete the course[39]. There was also a considerable amount

[39]I also note that most students adding the course after the initial registration period were more likely to end up in a traditional classroom. One explanation of this besides an *ex ante* preference for the traditional classroom is a more flexible capacity constraint in the traditional classrooms.

	Destination						
	A		B		C		
	Trad.	Active	Trad.	Active	Trad.	Active	Dropped
A Trad.	68	1	3	1	6	2	19
A Active	22	29	12	0	6	8	24
B Trad.	4	0	58	2	2	1	34
B Active	3	0	14	51	5	3	24
C Trad.	4	0	5	2	43	7	38
C Active	10	0	8	2	15	38	28
Added	26	4	24	7	26	13	

Table 3.2: Transitions from Initial Assignment (%)

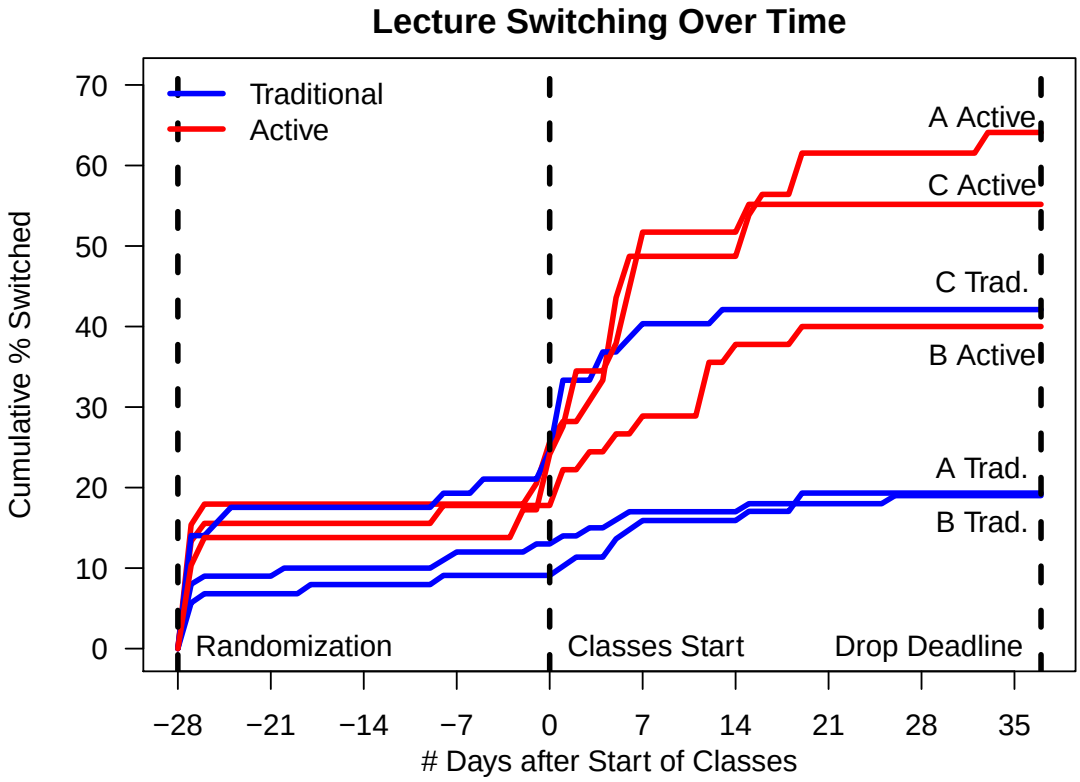

Figure 3.2: Diffusion of Students from Initial Assignment over Time

of noncompliance with treatment status, as many students switched from their initially assigned pedagogical style to the alternative; this was much more common for students changing from active to traditional classrooms, especially in time slots A and B.

Figure 3.2 presents the cumulative hazard curves for each lecture section, differentiated by pedagogical type. Very few students changed section before the start of classes; the exodus of students from initial assignment is largely concentrated in

Herein I treat these non-experimental students as I do the upperclassmen and exclude them from the analysis.

the first two weeks of "treatment exposure" (i.e., of attending their assigned class)[40]. The active sections in Time Slots A and C saw an explosion of departures in the first few weeks of the semester – more than 60% of initially assigned students would ultimately leave these sections. As seen in Table 3.2, very few people who switched sections opted to join an active learning section.

Students expressed a clear revealed preference for the instructors of Time Slots A and B. *Ex ante*, this could have been anticipated, as a quick search of the university's instructor review system (openly available to all students and common cultural knowledge) shows a long history of high-quantile reviews for both the A and the B slot traditional instructors. By contrast, only Time B's active instructor has recorded experience teaching the same course at the university prior to Fall 2015. As seen in Figure 3.2, this could be what is reflected in just how quickly students abandoned their initial assignments – risk-averse first-year students are faced with a culture shock in adjusting to college life and flock to the veteran instructors as a source of stability and reliability.

[40]I currently lack demographics for students not present in the calculus course at end of semester, and are thus unable to undertake an exercise of trying to find the best predictors of course dropping. I did explore predicting whether students switch pedagogical modes, but there are no strong predictors of either direction of switching among the individual-level observables I obtained. The strongest predictor of changing lecture sections is initial assignment, suggesting some instructor fixed effects which were confirmed anecdotally.

	Changed Treatment Status
Pell Eligible	0.12
	(0.07)
SAT Writing (100)	0.08
	(0.04)
Num. obs.	425

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.3: Using Observables to Anticipate Attrition

Predicting Switching

To find the best predictors of lecture switching, I used forward stepwise selection to pick from a list of covariates to minimize the cross-validated (ten-fold) misclassification rate; this process identified the best predictors in my data of switching lectures to be the student's math diagnostic score and Pell grant eligibility. The average misclassification rate for this triplet of covariates was 43, which is not all that much better than random guessing; the weakness of these predictors is confirmed in the summary of this regression in Table 3.3.

Variable	Overall	Active	Traditional	p-value
% Male	50	46	52	0.29
Ethnicity (%)				
White	40	32	43	0.27
Black	9	11	9	
Hispanic	12	14	11	
Asian	17	21	16	
Other	22	21	22	
Age	17	17	17	0.59
U.S. Citizenship (%)				
US Citizen	81	81	81	0.99
Perm. Res.	6	5	6	
Non-Res. Alien	13	13	13	
HS GPA	3.9	3.9	3.9	0.43
% without HS GPA	1	1	1	
% from US	91	91	92	0.74
U.S. Region (%)				
Mid-Atlantic	47	46	48	0.86
New England	4	4	4	
Midwest	10	13	9	
South	13	14	13	
West	10	9	10	
Rest of World	15	14	16	
SAT (of 2400)	2100	2100	2100	0.91
% Missing SAT	19	21	18	
ACT	33	33	33	0.89
% Missing ACT	50	49	50	
First-Year Need				
Average ($1000)	25	24	26	0.58
Median ($1000)	2	0	15	0.64
% Pell Eligible	14	16	13	
Program (%)				
College	77	78	76	0.80
Engineering	9	9	9	
Nursing	1	0	1	
Wharton	14	13	14	
# Observations	360	110	240	

Table 3.4: Descriptive Statistics by Initial Assignment

Analysis

Covariate Balance

I demonstrate the adequacy of random assignment in Table 3.4. All categorical variables are tested for independence of assignment with a Pearson's χ^2 test; continuous covariates are tested with a two-sample t test against the two-sided no-mean-difference null hypothesis. The median first-year need is tested with a Wilcoxon rank sum test, which has the null hypothesis that the distributions of gross need differ by a location shift of 0. Figures have been rounded to help preserve anonymity, so percentages may not sum exactly to 100. Balance is excellent, with only the highly skewed covariates (such as HS GPA) having p-values below .8, so I are confident that randomization was carried out successfully.

Evaluation Framework

The Rubin potential outcomes model (Rubin 1974) is the traditional workhorse of causal econometrics in situations like that at hand where establishing a causal relationship is not trivial – despite the considerable effort exerted in designing an orthogonal covariate, treatment attrition means I still must take care in specifying what exactly I're identifying.

Letting $Y_i(0) = \alpha_0 + \varepsilon_i(0)$ be an individual i's final exam score if they take the traditional version of the course and $Y_i(1) = \alpha_0 + \alpha_1 + \varepsilon_i(1)$ be the score for the

same student if they take the active learning version of the course, I can express each individual's observed score (noting that only one of $Y_i(0)$ and $Y_i(1)$ can ever be observed) Y_i as:

$$\begin{aligned} Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\ &= \alpha_0 + \alpha_1 D_i + \varepsilon_i(0) + (\varepsilon_i(1) - \varepsilon_i(0)) D_i \end{aligned} \tag{3.4.1}$$

Where D_i is an indicator taking the value 1 if student i experiences the active learning course and α_1 is the treatment effect. If receipt of active learning instruction were assigned randomly, the error term $\nu_i = \varepsilon_i(0) + (\varepsilon_i(1) - \varepsilon_i(0))D_i$ would be orthogonal to D_i (since $\varepsilon_i(1)$ can be constructed to be mean-0) and the treatment effect could be estimated with simple OLS (the point estimate being simply the difference in average test performance between the treated and untreated groups).

Alas, it is not receipt of instruction but lecture assignment that is random in my setup. As a result, D_i is likely to be correlated with $\varepsilon_i(1) - \varepsilon_i(0)$ – in a simple discrete choice framework, $D_i = 1 \Leftrightarrow \varepsilon_i(1) - \varepsilon_i(0) \geq -\alpha_1$, i.e., selection on the unobservables $\varepsilon_i(s)$ mean a student's choice to switch out of the active learning section signals their predilection to benefit from active instruction.

All is not lost, however, since I still have an orthogonal piece of information which will help separate selection effects from causal ones. First, I define T_i as an indicator for random assignment to an active learning classroom. The **intent to treat** (ITT) parameter is then given by β_1 in:

$$Y_i = \beta_0 + \beta_1 T_i + u_i$$

That is, the ITT is the average difference in performance between those assigned to treatment and those not,

$$\beta_1 = \mathbb{E}[Y_i|T_i = 1] - \mathbb{E}[Y_i|T_i = 0]$$

Typically, one would use OLS regression to estimate this and get its standard errors, but the relatively small sample and blocked nature of randomization lead us to prefer a permutation test to evaluate the statistical properties of my point estimate. Secifically, the permutation test (which I also run within each time slot to evaluate effect heterogeneity) consists of permuting the assigned treatment within each time slot and calculating the difference in average raw exam scores that would have obtained under this alternative treatment assignment, then repeating this dummy-assignment process 10,000 times. Under the null hypothesis that there is no difference in performance between treatment and control, this process will generate a distribution of reasonable estimates that may have occurred just due to sampling variability in the population at hand. The test concludes by using this distribution to calculate the p-value which is the percentage of permutations under which the observed difference in means is at least as extreme as that actually observed in the data. If the null hypothesis was assumed incorrectly, the observed difference in raw scores should be unlikely to have occurred in a permutation of group assignments.

Finally, I provide local average treatment effect estimates. The LATE approach is to consider D_i's linear relationship to T_i:

$$D_i = \gamma_0 + \gamma_1 T_i + \xi_i$$

Combining this with Equation 3.4.1 produces an instrumental variables framework with this as the first stage, i.e., T_i being randomly assigned ensures it is a valid instrument for D_i. Thus, α_1 can be estimated even in the presence of the endogeneity of D_i; given the binary nature of both D_i and T_i, it is possible to express this estimate in terms of expectations (this is often referred to as the **Wald Estimator** after Wald 1940):

$$\alpha_1 = \frac{\mathbb{E}[Y_i|T_i = 1] - \mathbb{E}[Y_i|T_i = 0]}{\mathbb{E}[D_i|T_i = 1] - \mathbb{E}[D_i|T_i = 0]}$$

The LATE gives the treatment effect for compliers – those that were induced to take the active learning version of the course by having been assigned. I estimate it as well with a permutation test approach.

Results

Table 3.5 presents the overall letter grade distribution, as well as its breakdown by treatment status. The differences appear most striking in the composition of students between the A and B letter grades for the two pedagogical styles, with more treated students earning a B and more control students earning an A. The

	Overall	Active	Traditional
A	42	38	43
B	35	38	34
C	17	15	18
D	2	4	2
F	0	1	0
W	3	4	3
Avg. GPA	3.2	3.1	3.2
Final Score	10.0	9.7	10.1

Table 3.5: Outcome Data by Initial Assignment

"Average GPA" row quantifies this more numerically by converting each student's letter grade assignment to the corresponding grade point average and averaging this across students in each treatment status, and confirms that control students have a higher overall average GPA (I turn to the statistical properties of this point estimate momentarily).

One drawback of using the final letter grade is its reflection of instructor feedback from converting each students' amalgamation of grades on homework assignments, the midterm, and the final into a letter grade. This mapping has subjective elements which may differ among professors or among treatment statuses (e.g., active classroom professors may be more or less lenient to their students given the more regular and intimate interaction they have due to the nature of their approach).

	All Students	Full Compliers	Partial Compliers
Intercept	0.15^{*}	0.12	0.13
	(0.07)	(0.07)	(0.07)
Active Learning	-0.35^{*}	-0.41^{*}	-0.34^{*}
	(0.15)	(0.19)	(0.17)
Num. obs.	345	254	287

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.6: Average Effects of Treatment on the Treated

To counteract this and get a more reliable measurement of student learning in the course, I use as my main outcome of interest from the course the unadjusted raw score on the common final exam. Every student in the course takes the same exam at the same time, and the panoply of instructors and teaching assistants comes together to grade all exams. These raw scores are then curved once the distribution of scores is known, and this is finally fed into each student's portfolio of grades and their final grade assigned. The raw score is simply a number of correct/incorrect answers on the exam, and as such is the most objective measure possible of each student's achievement in the course. The average (out of 15) of the raw score is presented in the final row of Table 3.5.

	All Students	Full Compliers	Partial Compliers
Intercept	-2.15^{*}	-2.61^{**}	-2.81^{**}
	(0.88)	(1.01)	(0.90)
Active Learning	-0.33^{*}	-0.34^{*}	-0.29^{*}
	(0.13)	(0.16)	(0.14)
Male	0.10	0.15	0.17
	(0.10)	(0.11)	(0.10)
Black	-0.69^{***}	-0.87^{***}	-0.79^{***}
	(0.17)	(0.17)	(0.17)
SAT Score (100s)	0.11^{**}	0.13^{**}	0.14^{**}
	(0.04)	(0.05)	(0.04)
Num. obs.	345	254	287

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Coefficient on an indicator of student program was excluded for confidentiality.

Table 3.7: Average Effects of Treatment on the Treated (with Controls)

Treatment on the Treated

Tables 3.6 and 3.7 present results for three specifications measuring the average effects of treatment on students that ultimately completed the coure under an active learning framework, i.e., treatment on the treated. The latter table includes The scores were first scaled to facilitate interpretation of the values. Treated students perform worse in the course than do those who experienced the traditional mode of instruction[41]. The first column in both tables considers all students completing the course; the second column examines only students who conformed exactly to their initial assignment. The final column considers any student who did not change treatment status, i.e., who started and ended the course in an active learning environment (but perhaps changed instructors). All columns use recitation-level clustered bootstrap standard errors. These results show treated students pretty consistently performed about 3/10 of a standard deviation worse than non-treated students.

	ITT I	ITT II	ITT III	ITT IV
Intercept	0.11	-2.75^{***}	-2.52^{**}	-3.03^{***}
	(0.07)	(0.82)	(0.83)	(0.90)
Active Learning (Assigned)	-0.15	-0.12	-0.01	0.07
	(0.12)	(0.11)	(0.11)	(0.11)
SAT Score (100s)		0.12^{**}	0.08	0.08
		(0.04)	(0.04)	(0.04)
4.0 GPA in High School		0.26	0.24	0.24
		(0.14)	(0.14)	(0.13)
Math Diagnostic Score			0.06^{**}	0.06^{**}
			(0.02)	(0.02)
Assigned Instructor Rating				0.14
				(0.10)
Num. obs.	345	341	316	316

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Errors are from 10,000 section-clustered bootstrap replications. Sample sizes differ due to missingness of high school GPA and/or pre-course diagnostic score for some students.

Table 3.8: Regression-Based Intent to Treat

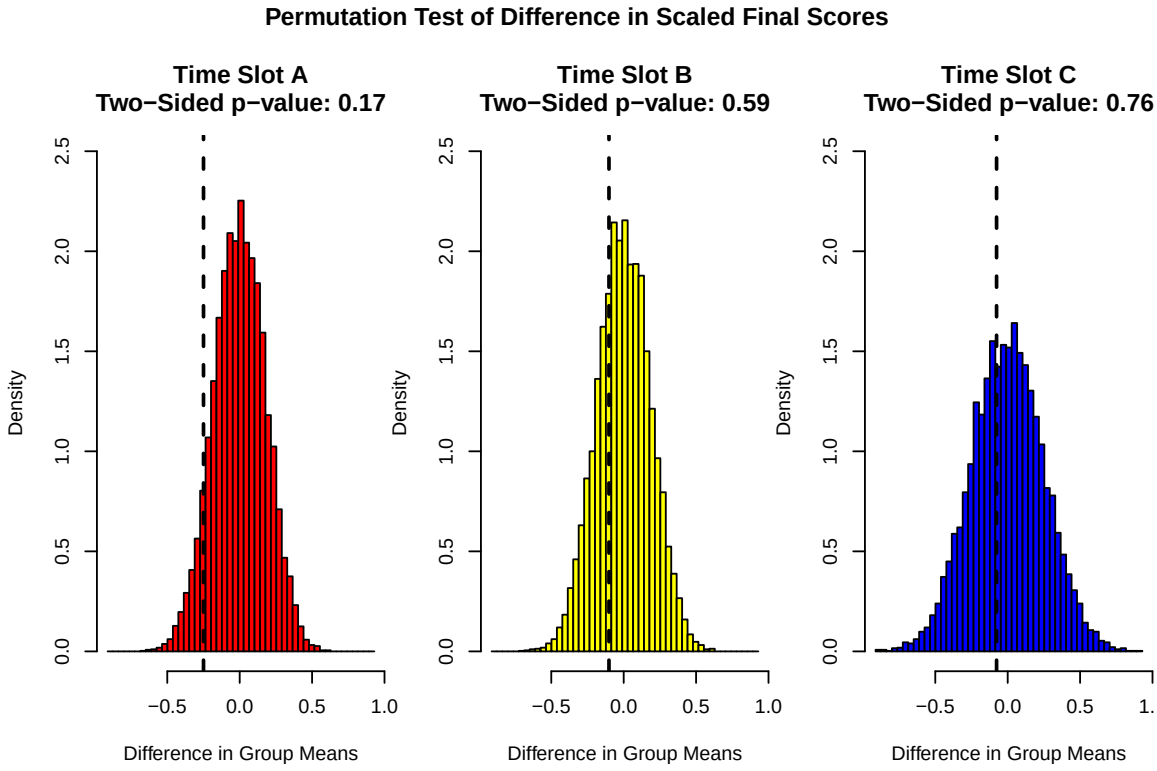

Figure 3.3: Permutation Distribution of ITT by Time Slot

	ITT I	ITT II	ITT III	ITT IV
Intercept	0.11	-2.75^{**}	-2.52^{**}	-3.03^{**}
	(0.06)	(0.83)	(0.86)	(0.94)
Active Learning (Assigned)	-0.15	-0.12	-0.01	0.07
	(0.11)	(0.11)	(0.12)	(0.13)
SAT Score (100s)		0.12^{**}	0.08	0.08^{*}
		(0.04)	(0.04)	(0.04)
4.0 GPA in High School		0.26^{*}	0.24^{*}	0.24^{*}
		(0.12)	(0.12)	(0.12)
Math Diagnostic Score			0.06^{**}	0.06^{**}
			(0.02)	(0.02)
Assigned Instructor Rating				0.14
				(0.11)
Num. obs.	345	341	316	316

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.9: Regression-Based Intent to Treat (Unadjusted SEs)

	ITT I	ITT II	ITT III	ITT IV
Intercept	0.11	-2.75^{***}	-2.52^{**}	-3.03^{**}
	(0.06)	(0.81)	(0.85)	(0.93)
Active Learning (Assigned)	-0.15	-0.12	-0.01	0.07
	(0.12)	(0.11)	(0.12)	(0.14)
SAT Score (100s)		0.12^{**}	0.08^{*}	0.08^{*}
		(0.04)	(0.04)	(0.04)
4.0 GPA in High School		0.26^{*}	0.24^{*}	0.24^{*}
		(0.12)	(0.12)	(0.12)
Math Diagnostic Score			0.06^{**}	0.06^{**}
			(0.02)	(0.02)
Assigned Instructor Rating				0.14
				(0.12)
Num. obs.	345	341	316	316

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.10: Regression-Based Intent to Treat (Asymptotic SEs)

Intent to Treat

Table 3.8 shows some regression-based estimates of the intent-to-treat effect[42]. The first column does a plain regression on the treatment indicator; the second and third columns add covariates likely to explain exam scores for precision[43]. The final column also includes a measure instructor quality for the instructor of the section to which the student was assigned. This measure is a 0-5 rating derived from student feedback. In none of the specifications is the initial assignment significant, i.e., the intent to treat effect is statistically zero.

To explain why (despite this difference being insignificant) the coefficient on as-

[41]Recall from the discussion above that, due to selection on unobservables, these estimates are likely flawed, as they are not immune to bias/inconsistency induced by student behavior. These results are presented for completeness, and the causes of the differences between these biased results and the unbiased results presented below are worthy of more detailed inquiry.

[42]Table 3.8 presents standard errors which result from 10,000 replications of resampling class sections and re-estimating the model. For completeness, Table 3.9 uses un-adjusted standard errors (i.e., the normal homoskedastic standard errors with $\mathbb{V}[\hat{\beta}] = \sigma^2(X^TX)^{-1}$), and Table 3.10 provides standard errors derived from asymptotic results. None of the standard error specifications affect the conclusions of the analysis. Though not strictly necessary to do the bootstrapped analysis, this approach can be justified given the skewed (non-normal) nature of the final exam score distribution – e.g., the p-value of the Shapiro-Wilk test for this variable is approximately 0.

[43]Of course, by including these variables, I have *a priori* reason to expect these may explain some of the variation in performance. Their significance does not indicate an issue with randomization (as a correlation of these characteristics with random assignment would), but rather their general importance as a proxy for latent student ability. The same applies to the LATE analysis below.

	ITT I	ITT II	ITT III	ITT IV
Intercept	0.11	-2.38^{**}	-2.52^{**}	-3.03^{***}
	(0.07)	(0.82)	(0.83)	(0.90)
Active Learning (Assigned)	-0.02	-0.02	-0.01	0.07
	(0.11)	(0.11)	(0.11)	(0.11)
SAT Score (100s)		0.11^{**}	0.08	0.08
		(0.04)	(0.04)	(0.04)
4.0 GPA in High School		0.27	0.24	0.24
		(0.14)	(0.14)	(0.13)
Math Diagnostic Score			0.06^{**}	0.06^{**}
			(0.02)	(0.02)
Assigned Instructor Rating				0.14
				(0.10)
Num. obs.	316	316	316	316

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.11: Regression-Based Intent to Treat (with Consistent Population)

	ITT I	ITT II	ITT III	ITT IV
Intercept	0.11	-2.65^{**}	-2.78^{***}	-3.32^{***}
	(0.07)	(0.80)	(0.82)	(0.89)
Active Learning (Assigned)	-0.15	-0.14	-0.14	-0.05
	(0.12)	(0.11)	(0.11)	(0.12)
SAT Score (100s)		0.12^{**}	0.10^{*}	0.10^{*}
		(0.04)	(0.04)	(0.04)
4.0 GPA in High School		0.25	0.22	0.22
		(0.14)	(0.14)	(0.14)
Math Diagnostic Score			0.05^{*}	0.06^{**}
			(0.02)	(0.02)
Assigned Instructor Rating				0.15
				(0.10)
Num. obs.	345	345	345	345

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Students missing HS GPA were assigned a value of 0 for having a 4.0 as the average SAT score for this group falls far below that of the average for either the 4.0 or the below-4.0 group. Math diagnostic scores were assigned by predicting their value from an OLS specification using only total SAT score.

Table 3.12: Regression-Based Intent to Treat (Using Imputed Covariates)

signment to active learning changes so dramatically from the second to the third specification, Table 3.11 and Table 3.12 provide estimations of similar models with slight adjustments to be sure the population in consideration is consistent. Table 3.11 does so by dropping students from Models I and II who are later dropped in Model III. Table 3.12 does so by imputing the GPA and/or diagnostic score from the student's SAT score, so that the population in Models II, III and IV matches that of Model I. Both make it clear that the source of the difference can be entirely ascribed to the poor course performance of students missing a diagnostic score.

Model IV also evinces a notable change in the magnitude of the coefficient of interest. The direction of the change indicate that the presence of higher-quality instructors in traditional classrooms is at least partially the source of the negative point estimates of treatment effects found in earlier models.

Figure 3.3 depicts a permutation test of differences in raw final scores grouped by time slot. The test was blocked by time slot to account more accurately for the blocked structure of randomization, and to control for any potential selection of student types into time slots (for example, student athletes typically have tightly constrained time schedules and can only possibly attend one time slot, and are likely to differ systematically from the rest of the student population in other meaningful ways).

In no case is the difference between the active and traditional classrooms on the final exam significant, confirming the course-level regression results from Table 3.8.

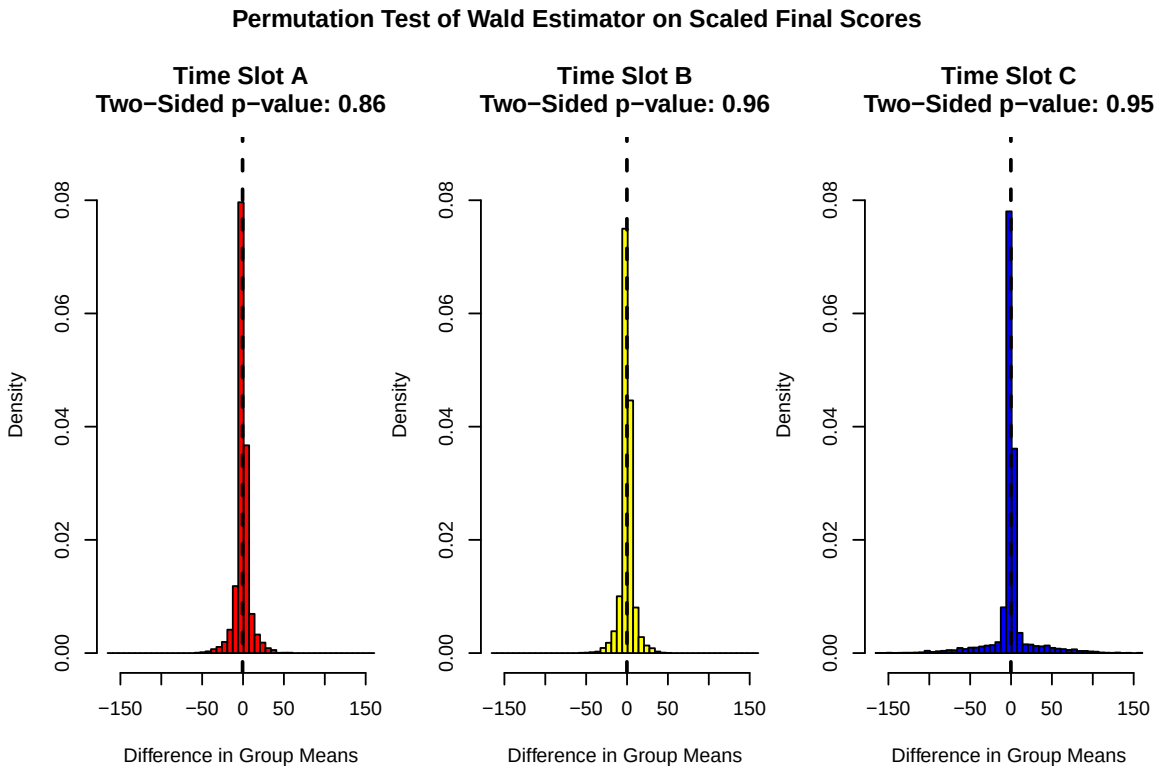

Figure 3.4: Permutation Distribution of Wald Estimator by Time Slot

The same goes for an aggregated version of the permutation test where the assignment is still permuted within time slots, but the difference in performance is calculated in the course as a whole (i.e., corresponding exactly to the first regression in Table 3.8). This point estimate is -0.15 and the associated p value is 0.2.

Local Average Treatment Effect

Tables 3.13 and 3.14 show models formulated with initial assignment as an instrument for treatment – hence, the coefficient on being in an active learning classroom corresponds to the local average treatment effect, namely, the treatment effect for

	LATE I	LATE II	LATE III	LATE IV
Intercept	0.07^{**}	-0.28	-0.41	-0.29
	(0.02)	(0.29)	(0.31)	(0.33)
Active Learning (Assigned)	0.52^{***}	0.52^{***}	0.46^{***}	0.44^{***}
	(0.04)	(0.04)	(0.04)	(0.05)
SAT Score (100s)		0.02	0.03	0.03
		(0.01)	(0.01)	(0.01)
4.0 GPA in High School		0.01	0.00	0.00
		(0.04)	(0.04)	(0.04)
Math Diagnostic Score			-0.01	-0.01
			(0.01)	(0.01)
Assigned Instructor Rating				-0.04
				(0.04)
R^2	0.33	0.32	0.28	0.29
Adj. R^2	0.32	0.32	0.27	0.27
Num. obs.	357	353	325	325
RMSE	0.35	0.35	0.34	0.34

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.13: First-Stage Regression for Local-Average Treatment Effects

	LATE I	LATE II	LATE III	LATE IV
Intercept	0.13	-2.81^{***}	-2.52^{**}	-3.01^{**}
	(0.07)	(0.82)	(0.86)	(0.94)
Active Learning (Assigned)	-0.27	-0.22	-0.01	0.15
	(0.21)	(0.21)	(0.24)	(0.29)
SAT Score (100s)		0.13^{***}	0.08	0.08
		(0.04)	(0.04)	(0.04)
4.0 GPA in High School		0.26^{*}	0.24^{*}	0.24^{*}
		(0.11)	(0.12)	(0.12)
Math Diagnostic Score			0.06^{**}	0.06^{**}
			(0.02)	(0.02)
Assigned Instructor Rating				0.15
				(0.11)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.14: Regression-Based Local Average Treatment Effects

	LATE I	LATE II	LATE III	LATE IV
Intercept	0.13	-2.81^{***}	-2.52^{**}	-3.01^{**}
	(0.07)	(0.80)	(0.85)	(0.92)
Active Learning (Assigned)	-0.27	-0.22	-0.01	0.15
	(0.21)	(0.21)	(0.24)	(0.30)
SAT Score (100s)		0.13^{***}	0.08^{*}	0.08
		(0.04)	(0.04)	(0.04)
4.0 GPA in High School		0.26^{*}	0.24^{*}	0.24^{*}
		(0.12)	(0.12)	(0.12)
Math Diagnostic Score			0.06^{**}	0.06^{**}
			(0.02)	(0.02)
Assigned Instructor Rating				0.15
				(0.12)
R^2	0.02	0.07	0.07	0.06
Adj. R^2	0.02	0.06	0.05	0.04
Num. obs.	345	341	316	316

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.15: Regression-Based Local Average Treatment Effects (Cluster-Robust SEs)

compliers. The first-stage estimates (of the endogenous treatment indicator vs. the exogenous random assignment indicator) are presented in Table 3.13, while the main coefficients of interest are in 3.14.

As expected given the earlier discussion of absence of strong predictors for non-compliance, only the instrument itself is strongly significant in the first-stage regression; this latter demonstrates that random assignment is indeed positively associated with receiving treatment. The main results, in Table 3.14, show that there is again no significant difference for treatment vs. control students. Table 3.15 repeats the same analysis with section-level cluster-robust standard errors and implies the same conclusion.

As above, the astute reader observed a large (though again insignificant) shift in the magnitude of the treatment coefficient from Models I and II to Model III. Again for robustness, Tables 3.16 and 3.17 account for this by using a consistent sample to estimate the models (the former) and by imputing missing values for incomplete observations (the latter). Again, I see that the suddent shift in coefficient is due to attenuated exam scores among the set of students without math diagnostic exam scores. A similar exposition as above applies to the change in coefficient resulting from including instructor quality as a regressor.

Figure 3.4 shows the permutation distribution of Wald point estimates generated alongside the ITT estimates produced for Figure 3.3. These distributions are characterized by very fat tails as the permuted difference in ultimate participation in the

	LATE I	LATE II	LATE III	LATE IV
Intercept	0.12	-2.39^{**}	-2.52^{**}	-3.01^{**}
	(0.07)	(0.87)	(0.86)	(0.94)
Active Learning (Assigned)	-0.04	-0.03	-0.01	0.15
	(0.25)	(0.24)	(0.24)	(0.29)
SAT Score (100s)		0.11^{**}	0.08	0.08
		(0.04)	(0.04)	(0.04)
4.0 GPA in High School		0.27^{*}	0.24^{*}	0.24^{*}
		(0.12)	(0.12)	(0.12)
Math Diagnostic Score			0.06^{**}	0.06^{**}
			(0.02)	(0.02)
Assigned Instructor Rating				0.15
				(0.11)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.16: Regression-Based Local Average Treatment Effects (Consistent Population)

	LATE I	LATE II	LATE III	LATE IV
Intercept	0.13	-2.72^{**}	-2.84^{***}	-3.33^{***}
	(0.07)	(0.82)	(0.82)	(0.90)
Active Learning (Assigned)	-0.27	-0.26	-0.25	-0.10
	(0.21)	(0.20)	(0.20)	(0.24)
SAT Score (100s)		0.13^{**}	0.10^{*}	0.10^{*}
		(0.04)	(0.04)	(0.04)
4.0 GPA in High School		0.24^{*}	0.22	0.22
		(0.11)	(0.11)	(0.11)
Math Diagnostic Score			0.05^{**}	0.06^{**}
			(0.02)	(0.02)
Assigned Instructor Rating				0.15
				(0.11)

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3.17: Regression-Based Local Average Treatment Effects (Imputed Covariates)

active classroom is often close to 0. Again the point estimates fail to show significance.

Chapter 4

Procrastination and Property Tax Compliance: Evidence from a Field Experiment

Taxpayers As Procrastinators

Most city residents are law abiding citizens. If late in their city tax payments it is unlikely it is part of a strategic plan to avoid ever paying. Property tax payments are computed by the city as assessed home value times the city's property tax rate and are known both to the city and the taxpayer. While it is possible to avoid payment by abandoning the property, this is very costly. For the vast majority of taxpayers the only issue is timely payment. Taxpayers receive their tax bill in January of the fiscal year with full payment or an agreed to payment schedule required by the end

of March. Most families have the payment withheld in an escrow account as part of their monthly mortgage payments. If payment, or enrollment in a payment plan, has not been made by the end of April, the city starts enforcement proceedings against the taxpayer. Enforcement begins with a reminder letter that all taxes and additional accrued interest and penalties are now due. In Philadelphia, those reminder letters are mailed in early May. I are studying the payment decisions of these tardy, or late, taxpayers. Following the analysis of O'Donoghue and Rabin (1999), my late taxpayer are seen as procrastinators who struggle with the problem of when, not if, to pay their property taxes.[44]

My taxpayer makes a decision every two weeks or perhaps every month as they pay their family bills. They can pay their taxes today, or postpone the decision until "tomorrow." If they pay their taxes today, they bear the immediate cost equal to the payment made. Taxpayers enjoy a benefit from having paid their taxes, but those benefits are not realized until "tomorrow," either as the simple relief of knowing their

[44]I are not the first to model taxpayer compliance as a problem of procrastination; see Hallsworth, et. al. (2014). I differ from their analysis in two ways. First, their focus is on late taxpayers as possibly credit-constrained households. That is less of an issue for my work as all my taxpayers are homeowners with assets that can used as collateral for a loan to pay taxes. It is true that homeowners, particularly the elderly, may not utilize such loans, but that is problem of financial literacy not tax compliance. Second, while I both rely upon the fundamental work of O'Donoghue and Rabin (1999), I amend that analysis to include the insight of Akerlof (1991) on the importance of "saliency" to the problem for procrastinators. I also extend the model to allow for active tax enforcement by the city.

taxes are paid or perhaps from the good feelings – that is, tax "morales" – of knowing they have met their obligations to their fellow residents.[45] This is O'Donoghue and Rabin's problem of the procrastinator facing immediate costs and delayed benefits. The decision period is today at time t, where t represents the number of periods since first receiving a notice that taxes are due. In deciding today as to whether to pay or not pay taxes, the taxpayer's inter-temporal utility function is specified over possible dates for payment. If the taxpayer makes a payment at time t, lifetime utility at time t is given by:

$$U_t^t = (\varphi^{t+1}\beta\delta)\, V - c_t \tag{4.1.1}$$

where c_t is the cost of tax payment at time t, and V is the benefit of knowing one's taxes are paid but not enjoyed until the period after payment. I assume V is constant for whenever taxes are paid.

Benefits are evaluated in today (period t) dollars allowing for declining saliency to future benefits and costs at rate φ ($0 \leq \varphi \leq 1$), possible present bias to all discounting at rate β ($0 \leq \beta \leq 1$), and the usual discounting of money values at rate δ ($0 \leq \delta \leq 1$). If the taxpayer plans to make a payment at time $t+s$, then the anticipated lifetime utility at time t of that payment is given by:

$$U_t^{t+s} = (\varphi^{t+s+1}\beta\delta^{s+1})\, V \; - \; (\varphi^{t+s}\beta\delta^{s})c_{t+s} \qquad s = 1, 2, ... \tag{4.1.2}$$

[45] I make the realistic assumption that my taxpayers will receive their public services whether they pay their taxes or not. If they do not pay their taxes, then they will be free-riding on the good will of their more responsible neighbors.

where c_{t+s} are the costs of tax payments at time $t+s$. The costs of tax payment may rise over time with accruing interest and penalties.

While tax payments made today are realized as a cost (c_t) today, tomorrow's tax payments and tomorrow's benefits are both realized in the next period, and are, therefore, discounted for today's decisions. Outcomes realized one period from today are discounted at the rate $\varphi^{t+1}\beta\delta$.

In my analysis the length of each individual period is relatively short, perhaps two weeks to a month between paying one's bills, and the overall decision horizon of my delinquent taxpayer's is no longer than several months. I will, therefore, assume that $\delta = 1$. The taxpayer may display a present bias, however, represented by a further discounting of future costs and benefits at a rate $\beta < 1$; time consistent taxpayers do not display a present bias so $\beta = 1$. Finally, my delinquent taxpayer may be forgetful which I represent as a declining rate of awareness or saliency, φ^{t+s}. Constrained by bounded rationality, taxpayers may only be able to pay attention to limited set of facts or tasks (Akerlof, 1991). For the forgetful taxpayer, $\varphi < 1$; for the fully aware taxpayer, $\varphi = 1$. In the extreme future or for the very forgetful taxpayer, $\varphi \simeq 0$ - that is, "out of sight, out of mind." Introducing the concept of saliency is a relatively simply way to give "reminders" an explicit role in taxpayer compliance.[46] As I discuss in detail below, saliency can explain differences in the response rates of taxpayers in the holdout sample and taxpayers that just received a neutral reminder letter.

[46] Saliency and reminders play a similar role in the behavioral economics of health policies; see Kessler and Zhang [2014] for a review.

My analysis focuses on the type of taxpayer who O'Donoghue and Rabin identify as the naive procrastinator. Here payment behavior stands in contrast to that of the fully aware ($\varphi = 1$) and time consistent ($\beta = 1$) taxpayer who will always pay her taxes on time (see below) and the sophisticated procrastinator who recognizes she is forgetful and/or present biased but is able to commit to an optimal payment schedule in advance. Here, that commitment device could be an escrow account with the mortgage bank or a city arranged tax payment plan. In contrast, the naive procrastinator assumes that she will remember to pay her taxes next period and do so in an optimal, time consistent way – but she does not. As a result, she may keep postponing payment until the end of the tax year when some drastic action – for example, court seizure of the home or garnishment of wages – is taken to collect all taxes, interest, and penalties due. Since both time consistent and sophisticated procrastinators will have paid, or have arranged to have paid, their property tax, they will not be in my sample of late taxpayers. Only naive procrastinators will be in my sample.

How does the naive procrastinator decide to pay her taxes? She will pay her taxes if the benefits from paying today are greater than benefits of paying at some later date. Following O'Donoghue and Rabin, I assume the naive taxpayer adopts what they call a *perception-perfect strategy* and pays her taxes today only if doing so gives them more perceived utility today than by paying at some future date. In my problem with constant V and rising costs c_{t+s} because of accumulating interests and

penalties, the best alternative date for paying taxes will always be in the immediate next period $t+1$. If so and assuming $\delta = 1$, the naive procrastinator pays today, if the lifetime utility of paying today is greater or equal to the lifetime utility if she delays:

$$(\varphi^{t+1}\beta)\, V - c_t \;\; \geq \;\; (\varphi^{t+2}\beta)\, V - (\varphi^{t+1}\beta)\, c_{t+1} \tag{4.1.3}$$

or if:

$$(\varphi^{t+1}\beta)\, (V(1-\varphi) + c_{t+1}) \;\; \geq \;\; c_t \tag{4.1.4}$$

The RHS of equation (4.1.4) is the perceived cost of paying one's taxes today. The LHS of equation (4.1.4) is the perceived cost of paying taxes one period later and is equal to the actual payment of those taxes one period later (c_{t+1}) plus the benefits "forgotten" ($V(1-\varphi)$) because of declining saliency. With time invariant benefits (V), if the perceived costs of paying one's taxes one period later are greater than or equal to the perceived costs of paying one's taxes today, the taxpayer will pay today.

Current period costs of compliance will equal taxes owed (T) plus accumulated interest and penalties at rate ρ now due from not paying taxes in prior periods:[47]

$$c_{t+s} \;\; = \;\; T\,(1+\rho)^{t+s} \quad s = 0, 1, 2, ...S, \tag{4.1.5}$$

[47]Strictly speaking interest and penalties do not begin to accumulate until some number of periods after the tax bill was first received. Rather than interest and penalties accumulating from the first date of the receipt of the tax bill for t periods as specified here, penalties only begin to accrue after a grace period. In the case of Philadelphia, the grace period between when the bill is received and taxes are due is three months. I adopt this simpler specification for the timing of payments to minimize the use of superscripts for dating all the periods. All that is required to ensure the same

Where S is the terminal date at which point a very large penalty is imposed upon the taxpayer for non-compliance, for example aggressive (harassing) enforcement or seizure of one's home. In the case of Philadelphia, after date S (December 31, 2015) the tax bill of the non-complying taxpayer, now called a "delinquent" taxpayer, can be given to a collection agency and the agency becomes the enforcer of payment. That agency can obtain a court-order to garnish wages of the violating taxpayer. As date S approaches the likelihood of compliance increases because of this very large, expected penalty.

Substituting this definition into equation (4.1.4) gives:

$$\varphi^{t+1}\beta\left(V(1-\varphi)+T\left(1+\rho\right)^{t+1}\right) \;\geq\; T\left(1+\rho\right)^{t} \tag{4.1.6}$$

as the requirement for current period tax compliance. More simply, rearrange and divide both sides by $T(1+\rho)^t$ and the condition for immediate tax payment becomes:

$$\varphi^{t+1}\beta\left(v(1-\varphi)+(1+\rho)\right) \;\geq\; 1 \tag{4.1.7}$$

where $v = V/[T(1+\rho)^t]$ are the benefits of paying one's taxes per dollar of taxes (and penalties) paid. The RHS of equation (4.1.7) is the cost of paying one dollar of taxes today; the LHS of equation (4.1.7) is the perceived costs of delaying and paying one's taxes in the next period. The perceived costs of delay are equal to the future benefits "forgotten" per dollar of taxes paid *plus* the added tax penalties from waiting. The

level of accumulated penalties is to lower the rate of interest and penalties, ρ, in my specification to reflect the grace period. All comparative statics from the model will be the same.

taxpayer will pay her taxes today if the cost of paying a tax dollar today is less than or equal to the costs of waiting and paying that tax dollar in the next period.

In contrast to the naive procrastinator who is forgetful ($\varphi < 1$) and/or present biased ($\beta < 1$) and may therefore delay payment, the fully aware ($\varphi = 1$) and time consistent ($\beta = 1$) taxpayer always pays her taxes on time – that is, with penalties and inerest, $1 + \rho > 1$.

In addition to the usual *passive* enforcement of late payments that occurs through the payment of interest and penalties when taxes are paid, the city may also use an *activist* enforcement strategy that audits some delinquent taxpayers at the beginning of the current period. If audited and determined to be a delinquent taxpayer, with probability π, the taxpayer must then pay an additional fine F in the next period. F might include "booting" the taxpayer's car, removing the taxpayer's children from school until payment is received, or additional fines equal to added administrative costs plus penalties. A city might target its activist strategy at those taxpayers with very large tax bills or with a year after year history of being a late taxpayer.

I assume, for simplicity, that activist enforcement is only in period t and not later.[48] If the taxpayer does not pay in period t, then under the activist enforcement strategy, the expected lifetime utility in the next period if there is delay must allow for the possible imposition of the penalty, F. In this case, the expected lifetime utility

[48]The extension to a model in which enforcement occurs in each period with probability π is not difficult and all results summarized in Proposition 1 also apply in that model.

from a one period delay becomes:

$$\begin{aligned} U_t^{t+1} &= \pi\left[\varphi^{t+2}\beta\, V - \varphi^{t+1}\beta c_{t+1} - \varphi^{t+1}\beta F\right) \;+\; (1-\pi)\left[\varphi^{t+2}\beta V - \varphi^{t+1}\beta c_{t+1}\right], \quad \text{or,} \\ &= \varphi^{t+1}\beta\left[\varphi V - c_{t+1} - \pi F\right] \end{aligned} \tag{4.1.8}$$

Now the taxpayer's decision rule is to pay if the expected utility of delay is less than the expected utility of paying today, or with the normalization that $f = F/(T(1+\rho)^t)$, if:

$$\varphi^{t+1}\beta\left(v\,(1-\varphi) + (1+\rho) \;+\; \pi f\right) \;\geq\; 1 \tag{4.1.9}$$

Note that the likelihood of making tax payments increases in the activist enforcement parameters π and f. The following proposition summarizes the analysis above.[49]

Proposition 1. *Naive procrastinating taxpayers will pay their taxes today if their perceived expected lifetime utility of delaying payment is less than or equal to the lifetime utility of paying their taxes today, or as long as the costs from delay are greater than the costs of payment today. The likelihood of payment will increase as:*

1. *taxpayer present bias is reduced (β rises);*

2. *taxpayer saliency of future benefits and costs increases (φ rises);*

3. *the benefits or the tax morale from the act of tax payment increases (v rises);*

[49]Equivalently, Equation 9 can be re-written as $\varphi^{t+1}\beta(v(1-\varphi)+(1+\rho)) \geq 1 - \varphi^{t+1}\beta\pi f$, where $\varphi^{t+1}\beta\pi f$ can be interpreted as a "benefit" of early tax payment or "forgiveness" of tax penalties. Penalty forgiveness is a common strategy to encourage tax payment.

4. *the penalties upon late payment or the subjective perception of these penalties increase (ρ rises); and*

5. *activist enforcement probability (π) and the fines (f) increase.*

Proposition 1 provides the conceptual framework for the design of my field experiment and the interpretation of my empirical findings. I design an experiment to evaluate the importance of three competing theoretical explanations of non-compliance: lack of salience (Proposition 1.2), or lack of benefits or tax morale (Proposition 1.3), or lack of deterrence (Proposition 1.4). I do not test Proposition 1.1, and therefore implicitly assume that all tardy taxpayers suffer from a common rate of present bias. Finally, my experimental design for Philadelphia does not allow us to evaluate the importance of activist enforcement strategies (Proposition 1.5).

A Field Experiment

The research setting for the experiment is the City of Philadelphia for calendar year, 2015. Notices of property tax payments are sent on January 1, and the full balance of taxes are due by March 31. If payment has not been received by that date, or the taxpayer has not entered into a tax payment plan with the City, then taxes are considered tardy and interest and penalties begin to accrue. On April 1, the City's Department of Revenue (DoR) begins contacting all taxpayers with unpaid accounts, informing them of taxes due and accumulated interest and penalties for late payment. At this time, the City will normally send two-thirds of the tardy

accounts to outside collection agencies acting as co-counsel for the City. The outside collection agencies are reimbursed at the rate of six percent of all their tardy revenues collected by December 31. The remaining one-third of the tardy accounts remain with the DoR for collection. All accounts still tardy on December 31 are designated as "delinquent" and then assigned to new outside collection agencies. For the purposes of my experiment the City of Philadelphia agreed to delay sending tardy accounts to the collection agencies until August 15, 2015.

My experiment was implemented with those taxpayers newly tardy on March 31, 2015. Of the 579,828 properties in the city receiving 2015 tax bills, approximately 100,000 or 17 percent were late in payment as of April 1. Of these 100,000 properties, 27,264 still owed more than $10 as of May 15 and had not owed property taxes from prior years. My experiment excludes all chronically delinquent taxpayers who owed taxes from prior years. Of the 21,468 tardy taxpayers, 2,429 taxpayers owned more than one property. While all 21,468 taxpayers were included in my experiment, I focus my empirical work on the 19,333 taxpayers who owned only one property.[50]

My experiment began with the mailing of reminder letters in mid-June, 2015 and continued to December 31, 2015. Of the tardy taxpayers with a single property, 16,940 received a standard or experimental reminder letter and 2,088 taxpayers did not receive a reminder. This sample of 2,088 taxpayers became my "holdout" sample and the basis for identifying the importance of saliency in taxpaying behavior. To

[50] As a robustness check I repeated my empirical analysis for the full sample of and the results are identical those I report in Sections IV and V below.

ensure that my experiment was not contaminated by other treatments not under my control, the DoR agreed to postpone all other enforcement activities until August 15. In particular, the outside collection agencies were not allowed to begin their collection efforts until after that date. The likely earliest date that those efforts led to any contact with a taxpayer is September 1.

Each reminder letter was approved by City's DoR to ensure that it could be understood by a taxpayer with at least a fourth or fifth grade level of English reading comprehension. Each letter also provided contact information for assistance for non-English speaking taxpayers. Translation were available for a number of different languages.[51]

Each reminder letter in my experiment was drafted to identify the possible impact on taxpayer compliance of the key variables in equation from Proposition 1. I could not, however, measure the effect of either taxpayer present bias (β) because my sample was limited to tardy taxpayers only. I also cannot evaluate the direct impact of a more activist enforcement strategy (π, f) as the city had not adopted such a strategy in my sample year, 2015. I can identify the potential importance of taxpayer saliency (φ), tax morales as they impact the benefits of tax payment (v), and interest and penalties (ρ). For brevity I present here the important distinguishing feature of each letter.

[51]Templates of the "reminder only" and "lien" letters are attached in the appendix. The full template for the other letters are available as an online appendix.

Reminder-only: **My records indicate that you have a balance due of *balance.*** If you have already paid, thank you. If not, please pay now or contact us to arrange a payment plan. The fastest and easiest way to pay is online at www.phila.gov/pay. Paying by E-check only costs 35 cent – less than the cost of a stamp!

The reminder-only letter allows us to identify the potential importance of tax saliency to taxpayer compliance. From Proposition 1 my holdout sample has a rate of saliency of φ^{t+1} when evaluating future benefits and costs. But those receiving my reminder letter today have a rate of saliency when evaluating future benefits and costs of φ only. When saliency is important, future taxes and benefits will be more salient after the receipt of the reminder, thus increasing the likelihood of taxpayer compliance; that is $\varphi > \varphi^{t+1}$, for $\varphi < 1$. A higher rate of compliance among taxpayers receiving the reminder-only letter compared to those in the hold-out cohort identifies a separate role for saliency in taxpayer compliance.[52]

Reminder plus Tax Lien: Failure to pay your Real Estate Taxes may result in a tax lien on your property in an amount equal to your back taxes plus all penalties and interest. When your property is sold, those delinquent tax payments will be deducted from the sale price. By paying your taxes now, you can avoid these penalties and

[52]My experimental design can identify the presence of saliency by an increase in compliance for those receiving a reminder letter, but time staggered reminder letters at a two-week or monthly interval would be needed to identify the actual rate of saliency – that is, the value of φ. This was not possible within the time constraints imposed by DoR on my experiment.

interest. Properties near you in your neighborhood that have liens placed on them include: < List Three Properties and Sale Dates > **Pay your taxes now to avoid a lien being placed on your property. My records indicate that you have a balance due of *balance*.**

Reminder plus Lien and Sheriff's Sale: Failure to pay your Real Estate Taxes may result in the sale of your property by the City in order to collect back taxes. In the past year I have sold N properties in your neighborhood at a Sheriff's Sale. Included in these N properties are the following properties near you: <List Three Properties and Sale Dates> **Pay your taxes now to prevent the sale of your property. My records indicate that you have a balance due of *balance*.**

The reminder letter coupled with the threat of a lien, or a lien plus a sheriff's sale of the taxpayer's home, increase the expected interest and penalties to the costs of delay – that is, an increase in penalties (ρ). Both letters make clear that interest and penalties will be collected by listing neighborhood properties where these added enforcement measures have been implemented. A taxpayer lien for all interest and penalties will be collected at the future date of home sale, which may be a very large obligation if the home is sold significantly in the future. A lien coupled with a sheriff's sale may occur sooner and thus have lower accumulated interest and penalties, but the forced sale of one's home is likely to have very high psychic costs. Which of the two added penalties is larger, and therefore likely to have a stronger impact on compliance, will depend upon the circumstances of the individual tardy taxpayer.

However, both letters should increase compliance over the holdout cohort from the reminder effect on saliency and from the added expected penalty, and both letters should increase compliance over the reminder-only letter from the added expected penalty.

My final four reminder letters test for the potential role of "tax morale" motives for compliance. An appeal to a tax morale is meant to cue a possible benefit from having paid one's taxes, apart from the actual receipt of services those payments may make possible. In contrast to user fees, property tax payments are not tied to the citizen's receipt of particular services during my experimental period. In effect, each delinquent taxpayer is a free rider, and the appeal to a tax morale for payment is meant to overcome such self-interest. In my model of taxpayer compliance these higher motives are captured by v in Proposition 1, the morale benefits from paying per dollar of taxes, interest and penalties paid.

I test for the importance of four such motives: 1) the value of knowing one is a contributor to the immediate services of one's neighborhood, v_N; 2) the value of knowing one is a contributor to the wider services that benefit the city as a whole, v_C; 3) the value of knowing one is part of a collective effort with other taxpayers or "peers" in paying for city services, v_P; and 4) the value of knowing one has meet one's obligations as a citizen in a democracy, v_D. Each of these benefits may motivate taxpayer compliance, and my reminder letters are meant to trigger a possible recognition of the importance of each motive. Some tardy taxpayers may respond to one motive,

some to another, and perhaps others to none at all if the free-rider motive is decisive. The four tax morale reminder letters are:

Reminder Plus Appeal to Neighborhood Services: I want to remind you that your taxes pay for essential public services in *neighborhood name*, such as <List Two Local Amenities>, your local police officer, snow removal, street repairs, and trash collection. **Please pay your taxes to help the city provide these services in your neighborhood. My records indicate that you have a balance due of *balance.***

Reminder Plus Appeal to City-Wide Services: Your taxes pay for important services that make a city great. Your tax dollars are essential for ensuring all Philadelphia's children receive a quality education and all Philadelphians feel safe in their neighborhoods. **Please pay your taxes as soon as you can to help us pay for these important services. My records indicate that you have a balance due of *balance.***

Reminder Plus Appeal to Peer Behavior: You have not paid your Real Estate Taxes. Almost all of your neighbors pay their fair share: 9 out of 10 Philadelphians do so. **By failing to pay, you are abusing the good will of your Philadelphia neighbors. My records indicate that you have a balance due of *balance.***

Reminder Plus Appeal to Civic Duty: For democracy to work, all citizens need to pay their fair share of taxes for community services. **By failing to do so, you are not**

meeting your duty as a citizen of Philadelphia. My records indicate that you have a balance due of *balance*.

The morale benefits from knowing one has paid one's taxes equals a weighted average of these motivations (v) plus a possible additional weight (v_i) when one of the reminder letters reinforces or enhances the affected benefit from tax payment: $v + \sum_i \omega_i v_i$, where i = N, C, P, or D, and where $\omega_i = 1$ if a reminder letter is received targeting benefit i, and v_i is the additional weight given to that motivation. I take as evidence that an increase in tax morale increases the likelihood of tax compliance when a tax morale reminder letter increases the rate of compliance above that of those receiving a reminder-only letter. If none of the tax morale letters impact compliance above a reminder-only letter then, at least on the margin for paying the property tax, the free-rider motivation is decisive for tardy Philadelphia taxpayers. In this case, increased enforcement will need to appeal to reminders and penalties.

Randomization Procedure

Randomization took place in two stages. As a baseline control, I randomly removed 3,000 tardy properties from the possibility of receiving any reminder letter at all, representing 2,088 property owners. These taxpayers (N=2,088) became my holdout sample and allowed us to estimate the efficacy of simply communicating with the taxpayer after the date that taxes are due. I next grouped all remaining properties by owner and randomized all owners to treatments based on the total amount of

property taxes owed on all of their properties.

While the vast majority of properties in the city of Philadelphia are owned by those with just one property, approximately 10 percent of the properties are owned by individuals or firms that own multiple properties. Since I are interested in taxpayer compliance and not property compliance, I identified owners of multiple properties by their legal name and randomly assigned each owner to a treatment group.[53] Any tardy taxpayer holding multiple properties within each treatment group received the same letter for each of those properties. Given the high correlation between the propensity to pay taxes and total debt owed, randomization blocks were defined according to owner-level total debt to assure uniformity of samples along the dimension of debt owed. Each property assigned to receive a reminder letter was equally likely to receive each of the seven treatments. Since most tardy property owners own only one property, my main interest in this study will be households that only own one property in the city. Once I restrict attention to this sample, I have 16,940 taxpayers in the treatment group and 2,088 taxpayers in the holdout sample. The total sample size is 19,028.[54] Table 4.1 checks whether the treatment and holdout groups are balanced based on the two most important variables, taxes due and assessed property

[53]I lacked an objective identifier such as a social security. There is some possibility that two or more different owners have the same name, but inspection by the authors found this to be very rare. To the extent that it occurs, I consider this random noise to the experiment.

[54]I also trimmed the sample and excluded the 28 owners with highest total assessed property value due to large variance in debt owed among the largest delinquents. None of the findings reported in this chapter depend on this trimming.

value.

Table 4.1 shows that randomization was successful in the single property owner sample. The average debt owed by each owner was $1,287 in the treatment group and $1,233 in the holdout sample. The average assessed property value is $144,145 in the treatment group and $142,630 in the control group. As a further test of my randomization procedure, I also checked to see whether randomization achieved spatial uniformity throughout the geographic expanse of the city. As reported in Table 4.1 geographic balance was achieved.

Next I test whether randomization was successful among the seven experimental treatment groups. Table 4.1 shows the results for the unary owner sample. Overall, I find no evidence that would suggest any problems with randomization. Results for multiple property owners, which do not differ from results for unary property owners, are reported in Table A2 in the appendix.

Empirical Results

Table 4.2 presents my core results for the three month period of my experiment largely unaffected by the intervention of the two outside collection agencies hired by the City to begin their own enforcement efforts in September, 2015. I consider two distinct measures of tax compliance behavior. First, did the taxpayer make any contribution at all towards their tax bill; this is the *ever-paid* response. Second, did the taxpayer make a full payment of their tax bill; this is the *paid-in-full* response.

Table 4.1: Balance on Observables (Single Property Owners)

Variable	Holdout	Reminder	Lien	Sheriff	Neighborhood	Community	Peer	Duty	p-value
Amount Due (June)	$1,233	$1,256	$1,280	$1,315	$1,289	$1,290	$1,280	$1,299	0.92
Assessed Property Value	$142,630	$158,370	$130,642	$134,334	$159,079	$130,265	$130,936	$165,617	0.53
Region									0.67
Center City	109	111	109	115	118	105	114	129	
Northeast Philadelphia	352	427	383	370	397	399	427	394	
North Philadelphia	449	520	525	526	491	498	533	527	
Northwest Philadelphia	537	601	645	666	620	654	615	611	
South Philadelphia	210	211	253	239	242	234	241	248	
West Philadelphia	431	549	514	500	519	551	486	523	
# Owners	2,088	2,419	2,429	2,416	2,387	2,441	2,416	2,432	

p-values in rows 1-2 are F-test p-values from regressing each variable on treatment dummies. A χ^2 test was used for the geographic distribution.

The sample includes only the 19,028 taxpayers who own a single property.[55] For ease of interpretation, Table 4.2 presents OLS estimates for the linear probability model; logit estimates are available in Tables A3 and A4 in the appendix and are identical in significance and interpretation to the OLS results reported here.

The top line of Table 4.2 reports the mean rate of compliance of my holdout sample for *ever-paid* or *paid-in-full* one month from the starting date of the experiment (July 15) and for the three months to the ending date of the experiment (September 15). The rate of *ever-paid* compliance for taxpayers in the holdout sample rises from 30.5 percent after one month to 51.4 percent after three months; the rate of *paid-in-full* compliance for the holdout sample raises from 23.5 percent after one month to 40.8 percent after three months. The rising rate of compliance for the holdout sample without receipt of a reminder letter is explained within the O'Donoghue and Rabin [1999] procrastination model by the presence of a terminal date to payment (S=December 31) at which time large costs to non-compliance can be imposed (e.g., garnishing of wages, sale of the home, publishing of names in the Philadelphia Inquirer).

The next seven rows report the additional impact on compliance from my seven treatment letters: Reminder-only, Reminder/Lien, Reminder/Sheriff, Reminder/Neighborhood,

[55]I have repeated my analysis for the sample of taxpayers, including multi-property owning taxpayers. Results for the full sample are identical to those reported here for unary (single) property owners. I limited my reported results and discussion to the single property owner sample. For comparison, results for the sample with multiple property owners are reported in Appendix Tables A1 and A4.

Reminder/Community, Reminder/Peer, and Reminder/Duty. Receiving the reminder-only letter increases the rate of compliance after one month for an *ever-paid* tax payment by 3.8 percent above the holdout's rate of compliance and by 3.9 percent after three months. Both effects are statistically significant at the 99 percent level of confidence. These estimates for the reminder-only letter indicate the relative importance of saliency to taxpayer compliance behavior. My letter is particularly effective early in my experiment, where the pure effect of a reminder increases the rate of compliance after one month by approximately 12 percent (= 3.8/30.5). While receipt of the reminder letter is still effective after three months, its relative impact on compliance behavior is less, adding an additional 8 percent (= 3.9/51.4) to the rate of *ever-paid*. The same statistical significance and declining rate of impact on compliance is observed for the outcome, *paid-in-full*. Here the reminder-only letter increases the one month rate of compliance over the holdout sample by 2.2 percent on a mean rate of holdout compliance of 23.5 percent (9.4 percent improvement) and the three month rate of compliance over the holdout sample by 3.0 percent on a mean rate of 40.8 percent (7.4 percent improvement).

Adding a message to the reminder letter has a mixed impact on tax payer compliance. Table 2 reports the joint effects of receiving a reminder and a message. Of the six messages, only the reminder/lien and reminder/sheriff letters had a statistically robust *added* impact on compliance. After one month, the sample receiving the reminder/lien letter had an additional 9.0 percent rate of *ever-paid* compliance over

the holdout sample's compliance rate of 30.5 percent rate (30 percent improvement) and after three months, an additional 9.2 percent rate of *ever-paid* compliance over the holdout sample's compliance rate of 51.4 percent (18 percent improvement). The impact is statistically significant at the 99 percent level of confidence. The results for paid-in-full compliance for the reminder/lien letter are also quantitatively important and statistically significant, adding 5.6 additional compliance over the holdout sample's one month mean rate of 23.5 percent (24 percent improvement) and an additional 7.2 percent compliance to holdout sample's three month mean compliance rate of 40.8 percent (18 percent improvement). Comparable impacts are observed for the sample receiving the reminder/sheriff letter, where I observe a 24 percent (=7.4/30.5) improvement in the rate of ever-paid compliance after one month, a 17 percent (= 8.8/51.4) improvement in *ever-paid* compliance after three months, a 19 percent (= 4.5/23.5) improvement in *paid-in-full* compliance after one month, and an 17 percent (= 6.8/40.8) improvement in textitpaid-in-full compliance after three months.

No such consistent improvements in compliance above the reminder-only letter are observed for those receiving a reminder letter with a "tax morale" message. This is seen most clearly in Table 4.3 where I compare compliance in the reminder-only sample to that of the samples receiving one of the six message letters. In this comparison, both the reminder/lien and the reminder/sheriff letters stressing the penalties of noncompliance have statistically significant and quantitatively important additional impacts on compliance above reminder-only, both for the *ever-paid* and *paid-in-full*

outcomes and at the one month and three month intervals. The lien letter adds more than a 5 percent increase in the rate of compliance above the reminder-only letter for *ever-paid* and about 4 percent to the rate of compliance for *paid-in-full*. These effects represent a 10 to 15 percent improvement in the rates of compliance over those obtained with the reminder-only letter. The sheriff letter also offers a significant improvement over the reminder-only letter, though the effects are slightly lower than those obtained with the lien letter. Compliance rates for *ever-paid* increase by 3 to 5 percent and for *paid-in-full* by to 2 to 4 percent above those achieved with the simple reminder. These effects represent a 9 to 11 percent improvement in compliance performance over what had been obtained with a reminder only. Table 4.3 also shows most clearly the inability of the tax morale reminders to induce greater compliance from Philadelphia's tardy taxpayers. Among those reminders, only the neighborhood letter is ever statistically significant and its effect is negative (!) for those paying in full.[56]

[56]My results for both the positive impact of penalties and mixed effectiveness of tax morale messages are consistent with most of the current literature on "nudges" and tax compliance; see Hallsworth [2014] for a thorough review. In the interest of full disclosure, however, my pilot study Chirico et al. [2016] for this project did find a role for a community or duty letter in increasing compliance. The control group in the pilot study received a reminder-only letter. Three other groups received either a penalty letter, a community letter – your taxes pay for city schools, police services, and fire fighters – or a combined peer/duty letter – 9 out of 10 Philadelphians pay their taxes; paying your taxes is your duty. In my pilot the penalty letter had no additional effect on compliance over that of the reminder-only letter. The community letter increased the rate of compliance above the

It is worth speculating as to why my results here differ from those in my pilot study. First, the pilot was run on a much smaller sample (3,900 single property taxpayers) and thus the results were less precisely estimated. Second, and more importantly, the sample for the study included only taxpayers who had not yet paid by the middle of November, 2014 (the time of my pilot), and thus are very close to being what the City will classify as a "delinquent" taxpayer as those who have not paid by December 31 of the tax year. The sample therefore consisted of the "most-tardy" of tardy taxpayers. Of these "delinquent" taxpayers who did make a contribution in my pilot study, the contributions were typically only partial payments of $50 to $150, suggesting these households may be seriously cash constrained. One might then imagine that for this sample of tax payers penalties are irrelevant; they cannot pay in full in any case. But a morale nudge might induce some payment in the spirit of a "charitable contribution." Consistent with this possible explanation is the fact that the average rate of compliance of this sample over the six weeks of my pilot was only 15 percent and the moral nudges boosted the rate of those making even some contribution to no more than 20 percent. It would be very valuable to design a larger experiment that seeks a compliance strategy for these very tardy or delinquent taxpayers. For this sample, one could "cost" to early payment in Equation 9 to reflect a cashflow constraint there. Here the literature on liquidity constraints

reminder letter by 4 percent, but the effect was not quite statistically significant. The combined peer/duty letter increased rate of compliance above the simple reminder letter by 2 percent and the effect was statistically significant at a 95 percent level of confidence.

is relevant; see Zeldes [1989]. My results are similar in statistical significance and impact to those in Castro and Scartascini's [2015] study of property tax payments in Junin Argentina, the other major field experiment seeking to improve property tax collection. For Philadelphians at least, and for the residents of Junin, it is reminders and penalties that improve compliance among tardy taxpayers.

Table 4.4 estimates the longer run impacts of my seven nudge interventions on compliance. The letters were sent on June 15th and received soon thereafter. The first two columns of Table 4.4 show the estimated effects on compliance of having received a letter six months later, again compared to compliance behavior in my holdout sample. Now the reminder-only letter no longer has an impact on compliance behavior, suggests declining saliency over time. Reminder letters that stress penalties from a lien or a lien plus sheriff's sale still have influence, however. The implied increase in expected penalty from non-compliance appears sufficient to overcome the loss of saliency. But again consistent with declining saliency, the estimated impact of the lien and sheriff letters, while still statistically significant, are roughly half as large as their impact at the one and three month intervals; compare Tables 4.2 and 4.4. Again, none of the tax morale nudges show a statistically significant impact on compliance behaviors. [57]

[57]The six month results need to be interpreted with care, however, as they are no longer part of my experimental design. Beginning between mid-August and mid-September the City allowed two private collection agencies to begin their efforts at collecting taxes from those in my original sample of 19,333 tardy taxpayers who had not yet paid their taxes, including those in my holdout sample.

The last two columns of Table 4.4 carry my sample into the next tax year, beginning with the receipt of a new property tax bill in early January, 2016, and asks if having received a reminder letter in June, 2015 improves compliance behavior for the payment of the 2016 taxes by June of 2016. Consistent with the importance of saliency, none of the 2015 reminder letters appear to have "staying power" into the next tax year. Tardy Philadelphians need constant reminders.

Discussion

While of interest as a specification and test of a behavioral theory of tax compliance, my results are also relevant for city tax collection policies. As a strategy for improving collection from tardy taxpayers, my analysis informs two important policy issues. First, cities need revenues: Do reminders improve collection, and then do reminders with a message raise more money than a simple reminder? Second, in light of the recent municipal fiscal crises and the potential for an unraveling of citizen commitment to local governance: Do reminders with a message, and then which message, improve tax collection as a "nudge" to citizen engagement? Table 4.5 provides answers to these two questions.

Listed in Table 4.5 are my seven treatments, the sample size to which each treat-

The treatments therefore, become a joint intervention of my letters and the unspecified, proprietary strategies of the collection agencies, which I then compare to the collection agencies' strategies alone as they impact those in the holdout sample. Whatever impact those proprietary strategies may have on compliance, my lien and sheriff sale letters still appear to have a lingering, value-added impact.

ment applied and total taxes owed, and then estimates of the impact of each treatment on the number new payers three months after receipt of the treatment letter, the average new revenue received per letter sent, total new revenues collected from each treatment letter above that paid by the holdout sample, and finally, the percent of owed taxes paid because of each treatment.

For single property owners, the total number of new taxpayers above the holdout sample from all reminder letters is 838, an average increase in the overall rate of compliance from receiving one my treatment letters of 4.9 percent (838/16,940). Table 4.5 also provides an estimate of additional revenues raised by each of my treatment letters and then the total revenue raised from each treatment group. From the perspective of the City's Department of Revenue, my experiment was a good investment of Department resources. Each letter cost about $1 to process and send. Thus estimated benefit to cost ratios for the seven treatments ranged from a low of $19.77 (the Neighborhood letter) to a high of $67.67 (the Lien letter). The approximately $17,000 spent on my experiment to mail the 16,940 treatment letters raised $615,752 in additional city revenues: an average benefit to cost ratio of 36.3.

Among my seven treatments, my experimental results clearly show the power of the lien and sheriff letters compared to a simple reminder or the tax morale nudges. The number of new taxpayers above the holdout sample is three to four times larger and the revenue/letter is two to three times larger with the letters stressing penalties. As a consequence, total new revenues (above the holdout sample) from the penalty

letters and new revenues as a share of all taxes owed are three to four times larger as well. If I had sent only the lien or sheriff's letter to the 16,940 taxpayers in my treatment groups I would have raised $1.15 million in new revenues rather than $616,752 – nearly twice as much. The paid share of taxes owed would have risen from my experiment's average of .028 to lien letter only of .053.

While the seven treatments are effective on the margin and the penalty letters particularly so, the final column makes clear that at least in Philadelphia, my treatments will not completely solve the larger problem of unpaid City property taxes. The treatments encourage a 3 to 9 percent higher rate of compliance above the holdout sample, and the typical new taxpayer pays on average about 60 percent of what they owe. [58] Thus the contribution towards total taxes owed will range from a low of 1.5 percent for the neighborhood letter to a maximum of 5.3 percent for the lien letter. Nudges help, and money is money, but at least in Philadelphia, they alone will only partially solve the large problem of tardy and then delinquent tax payments.

Money may not be all that matters with tax collection, however. Voluntarily paying one's taxes on time is a signal that one believes in what government is trying to do; see Posner [2000]. From the U.S. Colonies' resistance to British taxation in the 1760's to the boycotts of the apartheid government's imposition of utility taxes on the residents of Soweto in the 1980's, refusing to pay one's taxes is a rejection of government's performance. In signaling games where there is a cost to non-compliance, the more

[58] The median taxpayer in my sample who pays taxes, pays $738 towards the (average) tax bill of about $1200, or 60 percent.

who indicate they favor your contrarian position, the more likely you are to publicly express that position too; see Lohmann [1994] and Benabou and Tirole [2011]. In my case, what may have once been a strong tax compliance outcome can unravel to a new, non-compliance equilibrium when government no longer performs as needed for a majority of citizens; see Besley, Jensen, and Persson [2015]. Recently, such an unraveling towards a low compliance equilibrium can be observed in Detroit. The city's rate taxpayer compliance for property tax collections fell from a ten year average of .90 from 2000-2010 to a compliance rate of .68 by 2014 (Chirico, et. al., 2015). In 2013, 47 percent of Detroit's properties were classified as delinquent.[59] While nudges help, a high initial value of V reflecting government benefits significantly greater than tax costs may be the most important determinant of the aggregate rate of taxpayer compliance and commitment to city government; see Haughwout, Inman, Craig and Luce [2004].

[59]See Reese and Sands [2013] who conclude from their review of the economic and political events leading to the Detroit fiscal crisis that "it is not surprising that many view the social contract between property taxpayers and city government as broken." (p. 9) Another example of this can be seen in the 1990 taxpayer revolt to Prime Minister Thatcher's introduction of a local poll (head) tax; see Besley, Jensen and Persson [2015]. The regressive poll tax replaced a proportional property tax. In response to widespread citizen resistance the poll tax was removed two years later and the property tax restored. But compliance rates for the restored property tax were 14 percent lower than before: .83 vs. .97. Efforts to restore compliance since then have stressed high penalties but it has taken nearly eighteen years to return to the original rates of payment. Expected penalties perhaps are no substitute for good governance for ensuring voluntary taxpayer compliance.

Table 4.2: Short-Term Linear Probability Model Estimates

	Ever Paid		Paid in Full	
	One Month	Three Months	One Month	Three Months
Holdout	30.5	51.4	23.5	40.8
Reminder	3.8***	3.9***	2.2*	3.0**
	(1.4)	(1.5)	(1.3)	(1.5)
Lien	9.0***	9.2***	5.6***	7.2***
	(1.4)	(1.5)	(1.3)	(1.5)
Sheriff	7.4***	8.8***	4.5***	6.8***
	(1.4)	(1.5)	(1.3)	(1.5)
Neighborhood	1.7	2.7*	−0.2	1.5
	(1.4)	(1.5)	(1.3)	(1.5)
Community	3.8***	2.8*	1.3	2.5*
	(1.4)	(1.5)	(1.3)	(1.5)
Peer	3.9***	3.5**	1.8	3.4**
	(1.4)	(1.5)	(1.3)	(1.5)
Duty	2.4*	3.6**	0.7	2.3
	(1.4)	(1.5)	(1.3)	(1.5)
Num. obs.	19028	19028	19028	19028

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Holdout values in levels; remaining figures relative to the holdout benchmark.

Table 4.3: Short-term Results: Relative to Reminder-Only

	Ever Paid		Paid in Full	
	One Month	Three Months	One Month	Three Months
Reminder	34.3	55.4	25.8	43.8
Lien	5.3***	5.3***	3.4***	4.2***
	(1.4)	(1.4)	(1.3)	(1.4)
Sheriff	3.6***	4.9***	2.3*	3.7***
	(1.4)	(1.4)	(1.3)	(1.4)
Neighborhood	−2.1	−1.2	−2.5*	−1.5
	(1.4)	(1.4)	(1.3)	(1.4)
Community	0.1	−1.1	−0.9	−0.5
	(1.4)	(1.4)	(1.3)	(1.4)
Peer	0.1	−0.4	−0.4	0.3
	(1.4)	(1.4)	(1.3)	(1.4)
Duty	−1.3	−0.3	−1.6	−0.7
	(1.4)	(1.4)	(1.3)	(1.4)
Num. obs.	16940	16940	16940	16940

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$. Reminder values in levels; remaining figures relative to this

Table 4.4: Long-Term Linear Probability Model Estimates

	Six Months		Subsequent Tax Cycle	
	Ever Paid	Paid in Full	Ever Paid	Paid in Full
Holdout	73.3	63.2	65.5	52.5
Reminder	1.3	1.5	−1.4	−0.7
	(1.3)	(1.4)	(1.4)	(1.5)
Lien	3.8***	4.8***	−0.9	−0.7
	(1.3)	(1.4)	(1.4)	(1.5)
Sheriff	3.8***	3.0**	−0.6	−1.1
	(1.3)	(1.4)	(1.4)	(1.5)
Neighborhood	−0.2	−0.0	−3.1**	−2.2
	(1.3)	(1.4)	(1.4)	(1.5)
Community	0.9	1.1	−1.8	−2.0
	(1.3)	(1.4)	(1.4)	(1.5)
Peer	1.3	2.3	−1.9	−1.1
	(1.3)	(1.4)	(1.4)	(1.5)
Duty	2.1	1.0	−1.6	−1.9
	(1.3)	(1.4)	(1.4)	(1.5)
Num. obs.	19028	19028	19025	19025

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Holdout values in levels; remaining figures relative to this

Table 4.5: Three Month Impact of Collection "Nudges"*

Treatment	Sample Size	Total Taxes Owed	New Payers	Revenue/ Letters	New Revenues	New % of Taxes Paid
Reminder	2,419	$3.038 M	95	$28.79	$69,643	.023
Lien	2,429	$3.109 M	224	$67.67	$164,370	.023
Sheriff	2,416	$3.177 M	213	$64.90	$156,798	.049
Neighborhood	2,387	$3.077 M	65	$19.77	$47,191	.015
Community	2,441	$3.149 M	68	$20.91	$51,041	.016
Peer	2,416	$3.092 M	85	$25.65	$61,970	.020
Duty	2,432	$3.159 M	88	$26.62	$64,739	.020
Totals	16,490	$22.143 M	838	-	$615,752	.028

* Sample Size are the number of single property taxpayers in the treatment group. Total Taxes Owed is the total taxes owed by single property taxpayers in the treatment group. New Payers equals the new payers after three months computed as the estimated increase in rate of compliance of those receiving the letter over those in the holdout sample as reported in Table 2; for example, for the reminder letter the number of new payers equals 95 = .039 x 2,419. Revenue per letter for each treatment equals the median new revenue collected from those who received a treatment letter and made some payment (=$ 738/letter) times the three month increase in compliance from each treatment letter; for example for the reminder letter the median estimated revenue per letter equals $28.79 = .039x$738. New revenues for each treatment equals the revenue/letter times the number of single owner properties receiving a treatment letter: for example, for the reminder letter the estimated total new revenues equals $69,643 = $28.79x2,419. New % of Taxes Paid equals New Revenues Divided by Total Taxes Owed; for example, for the reminder letter .023 = $69,643/$3,038,000.

Chapter 5

Conclusion

U.S. cities have a mixed record in their ability to collect taxes from their residents. Some (Boston, Charlotte, San Francisco, San Antonio) do a good job, collecting over 98 percent of property taxes due, others (New York, Philadelphia, St. Louis) are less successful, collecting in the neighborhood of 90 percent, and finally some, such as Detroit and Flint, collect less than 70 percent of property taxes owed (Chirico, et. al., 2016). Collecting property taxes should be straightforward; both the city and the property owner know exactly what is due. While scofflaws, those permanently in arrears, are a problem, most tardy tax payments are because residents forget or are hoping to "let it ride" and not be noticed. We provide in Chapter 3 an extension of the O'Donoghue-Rabin's 1999 theory of procrastination to explain this behavior. We then test three competing explanations incorporated in our model using a field experiment on property tax compliance in Philadelphia.

Our empirical analysis reached three conclusions. First, there is strong evidence that salience is important. A simple reminder will improve compliance. The rate of compliance rose by 4 percent with a simple reminder above that of our holdout sample that received no reminder. But the effects of the reminder decline over time. There is no evidence that having received a reminder in 2015, and having even paid your taxes, improves your chance of compliance when paying your 2016 taxes. These results strongly suggest tardy taxpayers lack salience which is consistent with Akerlof's 1991 work on procrastination. Taxpayers may have a limited capacity to remember and process tax (and benefit) information when making their spending and financial decisions. An explicit reminder that brings that information to the fore can encourage payment. In this regard our results are consistent with those in Chetty, Looney, and Kroft 2009 on the role of saliency in the payment of sales taxation and the results in Bhargava and Manoli 2015 on the take-up rate for welfare benefits.

Second, a reminder letter with a "message" can improve compliance above a simple reminder, but the content of the message matters. Two of our reminder letters stressed increased penalties for non-compliance; one threatened to place a lien on the property if taxes are not paid and the second threatened a lien and the risk of an immediate sheriff's sale if taxes are not paid. Both had a significant impact on compliance, raising the rate of compliance by 9 percent over that of taxpayers who received no reminder at all. This finding strongly supports the theory that tardy taxpayers lack sufficient deterrence. The messages that did not improve compliance above that of

a simple reminder letter were our four "tax morale" messages: one stressing your taxes are needed for neighborhood services such as trash collection and the local park, a second that your taxes pay for important city-wide services such as education and protection, a third that 9 of 10 other Philadelphians pay their taxes on time, and a fourth that paying one's taxes is an important component of the democratic contract. It is important to stress, however, that the impact of any nudge on behavior is conditional on the content of the message, its fiscal context, and affected taxpayers. Our results are for Philadelphia, *given* its current levels of penalties, the current level of services provided by the City, and the preferences of its tardy taxpayers. Tax nudges in cities with lower penalties, better services, or more civically minded taxpayers might induce different behavioral responses. That said, the similarity of our results, both qualitatively and quantitatively, to those of Castro and Scartascini 2015 for the property tax payments in Junin, Argentina is reassuring.

Third, the marginal impacts on city revenues of our strategies were quantitatively significant. A simple reminder letter earned the City $28 more in additional revenues for each additional dollar of administrative cost. A reminder coupled with our most effective messages - the tax lien and sheriff letters - earned the City $65 more in extra revenues for each dollar expended. This very high marginal revenue to cost ratio strongly suggests that well targeted nudges should be part of any City's revenue collection strategy; see Keen and Slemrod 2016.

The conclusions of the second chapter are much less stark. The concept of an

actively engaged classroom is one which appeals on some level to many math educators with whom we have spoken. And in fact many students have expressed a strong preference for the approach to tackling new material in mathematics that we analyzed in Chapter Two. Conversely, however, many students voiced their disdain for the active classroom and much preferred the traditional lecture format, which in some ways offers more flexibility. It should come as no surprise that the actively taught math classroom is not a silver bullet for bringing struggling students to mathematical enlightenment.

We have, however, given evidence to help confirm the active classroom's place in the versatile educator's toolbelt. In fact, it so happened that one of the "traditional" instructors (who has indeed is known pushing the bounds of what "traditional" means and was one of the early adopters of posting their recorded lectures for their students) decided to "activate" their lecture hall several times, segueing from a standard lecture to mid-sized groupwork mid-session; and their TAs were seen to run their recitation sessions in the "active" groupwork format.

In the end, the students performed slightly (though not significantly) worse on the course final exam, but even this taken alone would not amount to a condemnation of the utility of the active approach – after all, the more fundamental outcome of interest has not yet emerged, which is the ultimate persistence rates of active-assigned students in STEM fields.

Chapter One, too, came to less-than-monumental conclusions. That salary in-

centives appear to play such a limited role in driving teacher churn is bound at first glance to be a disappointment for policymakers. The most powerful predictors of turnover in educators in Wisconsin are all basically beyond the control of administrators, who have no readily-manipulated direct lever for assigning students to schools[60]. HKR found school quality (as measured by average standardized test performance) to be of key importance for attracting/retaining teachers, but we found no evidence that student proficiency (as measured by attainment levels on standardized tests) is a factor in the turnover decision for Wisconsin teachers. Regardless, manipulating school performance is famously difficult[61], and is in fact the original goal administrators often have in mind when they turn to labor market policies in the first place, so that telling administrators they can improve teacher retention by improving student performance amounts essentially to circular reasoning.

The upside is that this chapter is far from settling the debate about welfare-maximizing teacher turnover policies. Limitations in our data prevent us from associating to teachers anything but crude measures of their productivity; measures such as experience, certification, and race are famously poor predictors of teacher quality

[60]There is evidence (e.g., Richards 2014) that catchment area manipulation (educational gerrymandering) is being used by some schools to select their student populations, but the equilibrium outcome of the strategic interactions of districts competing for the most "desirable" students is far from clear.

[61]See, for example, the widespread cheating scandals on standardized tests by teachers in Chicago (Jacob and Levitt 2003), Atlanta, and Philadelphia as an example of the lengths professionals feel they need to go to effect change in testing outcomes.

measures such as value-added. We are thus unable to provide any input to the question of whether *high-quality* teachers have patterns of mobility which resemble that of the teaching population as a whole, or whether heterogeneity in their preferences can be used to devise appropriate policies.

Appendices

Appendix A

Appendix to Chapter 2: Longitudinal Teacher Panel from Unlinked Cross-Sectional Cuts

DPI WISEstaff data does not include a longitudinal identifier for teachers, so we resorted to an alternative approach to matching teachers from year to year. The essence of this algorithm relies on the inclusion of four fields in the DPI data – `first_name`, `last_name`, `nee` (former last name) and `birth_year`. By matching teacher using these identifiers, it is possible to construct with high accuracy a panel of teachers from simple teacher cross-sections[62].

[62]The code for this process was done using R and especially helped by the `data.table` package (Dowle and Srinivasan 2017). The code to reproduce this entire project can be found at `https://github.com/MichaelChirico/wisconsin_teachers`. The script for the algorithm is

Step 0: Pre-Processing

Prior to beginning the matching process, a number of steps are taken to improve the quality of the raw data. The first is to incorporate as many of the errata mentioned in Public Instruction (2017b) as possible. All name variables are then converted to lower case, after which we extract maiden names (identified for those missing a DPI-supplied entry in `nee` as the part of the `last_name` field that appears in between parentheses or surrounding a hyphen or forward slash). Generally, it appears the maiden name comes in the data *before* the hyphen, but we create the `nee2` field to identify potential matches to the post-hyphen name as well. A search was done of the data for irregular characters (punctuation or numbers) which allowed several obvious typos to be resolved (e.g., `l0is a dewey` was easy to resolve as being `lois a dewey`), and this is implemented next.

We then create a "clean" version of the name fields which strips away all whitespace, initials (lone letters), and punctuation. At this stage, all observations which identically match another in the same year from the viewpoint of the algorithm – namely, those that match exactly another observation on the cleaned first and last names and year of birth – are removed from the data since it would be impossible to tell such teachers apart. A more ambitious treatment would attempt to use other cues found in the duplicated records (ethnicity, subject/position cues, etc.) to separate

`teacher_match_and_clean.R`. The `README` file gives steps for full reproduction, including retrieving the raw data.

such teachers, but the marginal cost of doing so was found to exceed the potential benefit considerably for the exercise at hand (recall that only 0.7% of total observations are lost in this fashion). Finally, teachers in the first year of data (1994-95) are assigned an ID starting from 1 using the within-year identifier provided by DPI.

Steps 1-21: Matching

The algorithm proceeds by iterating over years of the data. In each year Y, matches are found serially by progressively adjusting the criteria for considering two observations to be from the same teacher as follows:

1. Match anyone who stayed in the same school – i.e., match any teacher found in a year $Y' < Y$ with the same `first_name`, `last_name`, and `birth_year` at the same `district` and `school`.
2. Find within-district switchers – those who match on all but the `school` field from Step 1.
3. Find out-of-district switchers – those who match on all but the `district` field from Step 2.
4. Find teachers that appear to have been married, but have not moved – their `nee` field in Y is matched to the `last_name` fields in $Y' < Y$, but otherwise the fields from Step 1 are all matched. We create a flag for teachers matched in this fashion called `married`.
5. Repeat Step 2 for those who appear to have married.

6. Repeat Step 3 for those who appear to have married. 7-9. Repeat Steps 4-6 using the `nee2` field instead of `nee`. 10-18. Repeat Steps 1-9 using the "cleaned" version of the first name field that had non-alphabetic characters removed, `first_name_clean`. We create a flag for teachers matched in this fashion called `mismatch_inits`.
7. Match individuals in the same school assigned to the same position (identified in `position_code`) but with different years of birth to overcome potential noise in year of birth (most commonly, the year of birth is missing in some years). We create a flag for teachers matched in this fashion called `mismatch_yob`.
8. Match individuals in the same *district* assigned to the same position but with different years of birth. We do not extend this logic to find district switchers since the potential for erroneous match assignment in such a case is too great, and we neglect to extend the algorithm to use other cues from the data to facilitate matching in such cases.
9. Assign new IDs to all teachers in Y not matched in the first 20 steps, incrementing from the highest ID recorded thus far.

To help ensure we are matching to the most important observation of each teacher, matching is always done to a teacher's highest-FTE observation within a year (particularly important for Steps 19-20). Further, it is sometimes the case that a given tuple of search keys matches more than one teacher in the prior data; if so, these rows are simply ignored for that step and such a teacher will go unmatched unless they

are uniquely pinned down in a subsequent step.

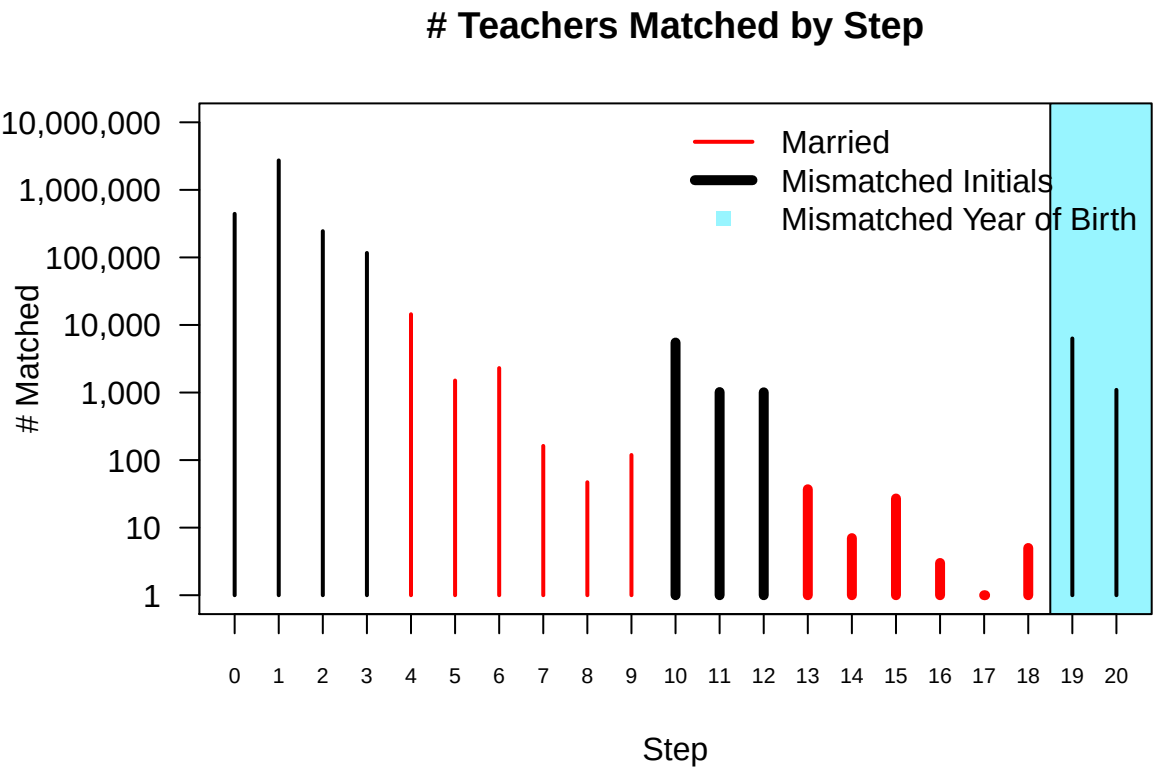

Figure A.1: Frequency of Matching by Step

Figure A.1 Shows that teachers are most commonly matched in the first three steps, meaning data fidelity issues are not *per se* devastating. The real benefits of the algorithm are in the subsequent steps, as a result of which an additional 33,538 teachers are matched than would have been if only the original first and last name fields as found in the raw data were used.

Step 22: Post-Processing

The panels implied by the matched IDs created by the algorithm still have a substantial amount of data quality issues which we can only address once teachers' multiple observations are associated, mainly having to do with instability in certain observable characteristics which should be constant over time. First, we cascade forward maiden names (if a teacher has non-missing `nee` in a period Y and it becomes missing in $Y' > Y$, we replace it in Y' with its value in Y); the same is done for the certification field, `highest_degree` (just as a teacher cannot erase marriage from their past, so can they not make a degree disappear).

Next, we correct instability in the `ethnicity` (and `gender`) fields when possible according to three steps: 1) it is sometimes missing, in which case we simply overwrite it with the other values found for that teacher; 2) at least 70% of a teacher's observations use the same ethnicity (or gender); or 3) there are at least five people that share a last name with an ethnicity-ambiguous teacher, at least 70% of whom have one ethnicity (or gender), the idea being that names like Xu or Gutierrez are strongly associated with a particular ethnicity. This type of correction is uncommon enough not to warrant an appeal to a more sophisticated approach commonly found in natural language processing applications, e.g. training a classifier such as a random forest (Breiman 2001) to predict ethnicity as accurately as possible.

Lastly, we synergize the year of birth field for those matched in Steps 19 or 20 by assigning the one that appears most frequently for each teacher; in the case of

ties, we use a regression-to-the-mean-type logic and assign the year which brings the teacher closer to the median age observed in the data. More data-driven approaches (conditioning the target median on the teacher's employer, position, year in the data, or using social security data to determine the maximum-likelihood year of birth for a given first name, etc.) were again eschewed for expediency.

Validity Check: `file_number`

Starting in the 2011-12 release, the DPI data begins to consistently record a field called `file_number` for teachers which generally acts as a time-consistent ID (from verbiage gleaned from Public Instruction 2017b, it appears this corresponds to a teaching license number). We looked for instances of multiple file numbers and are content that the algorithm is performing well – only 78 teacher IDs were found to be associated with more than one `file_number`, with almost all of them having been matched on Steps 1 - 3 (what should be the highest-accuracy steps). Given a number of apparent transcription mistakes (i.e., `file_number` differing by one digit in some years) and that the `file_number` does appear to change on occasion, even these 78 could be an overstatement of the number of incorrectly matched individuals.

Appendix B

Appendix to Chapter 4:

Additional Figures and Tables

The appendix contains Tables A2 and A1 which summarizes additional balance tests and robustness analyses using all owners (including multiple property owners).Tables A3 and A4 report estimates based on Logit models for unary owners and unary plus multiple owners.

Table A1: Robustness Analysis: Relative to Reminder (All Owners)

	Ever Paid		Paid in Full	
	One Month	Three Months	One Month	Three Months
Reminder	34.9	56.5	23.9	41.8
Lien	4.8***	4.7***	3.3***	4.0***
	(1.3)	(1.3)	(1.2)	(1.3)
Sheriff	3.4***	4.6***	2.3**	3.6***
	(1.3)	(1.3)	(1.2)	(1.3)
Neighborhood	−1.0	−0.8	−1.2	−0.4
	(1.3)	(1.3)	(1.2)	(1.3)
Community	−0.4	−1.4	−0.6	−0.2
	(1.3)	(1.3)	(1.2)	(1.3)
Peer	0.3	−0.8	0.4	0.8
	(1.3)	(1.3)	(1.2)	(1.3)
Duty	−1.3	−0.2	−1.0	−0.8
	(1.3)	(1.3)	(1.2)	(1.3)
Num. obs.	19333	19333	19333	19333

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Reminder values in levels; remaining figures relative to this

Table A2: Balance on Observables

Unary Owners

Variable	Reminder	Lien	Sheriff	Neighborhood	Community	Peer	Duty	p-value
Amount Due (June)	\$1,256	\$1,280	\$1,315	\$1,289	\$1,290	\$1,280	\$1,299	0.98
Assessed Property Value	\$158,370	\$130,642	\$134,334	\$159,079	\$130,265	\$130,936	\$165,617	0.46
# Owners	2,419	2,429	2,416	2,387	2,441	2,416	2,432	0.99

Unary and Multiple Owners

Variable	Reminder	Lien	Sheriff	Neighborhood	Community	Peer	Duty	p-value
Amount Due (June)	\$1,593	\$1,593	\$1,590	\$1,589	\$1,583	\$1,572	\$1,583	1
Assessed Property Value	\$180,664	\$155,499	\$157,398	\$180,172	\$153,528	\$155,438	\$183,991	0.48
% with Unary Owner	87.6	88.0	87.5	86.4	88.4	87.5	88.1	0.42
% Overlap with Holdout	3.69	3.44	3.29	3.73	3.40	3.55	3.40	0.97
# Properties per Owner	1.27	1.26	1.26	1.32	1.26	1.26	1.26	0.67
# Owners	2,762	2,761	2,762	2,762	2,762	2,762	2,762	1

p-values in rows 1-5 are F-test p-values from regressing each variable on treatment dummies. A χ^2 test was used for the count of owners.

Table A3: Short-Term Logistic Model Estimates (Unary Owners)

	Ever Paid		Paid in Full	
	One Month	Three Months	One Month	Three Months
Holdout	−0.8	0.1	−1.2	−0.4
Reminder	0.2***	0.2***	0.1*	0.1**
	(0.1)	(0.1)	(0.1)	(0.1)
Lien	0.4***	0.4***	0.3***	0.3***
	(0.1)	(0.1)	(0.1)	(0.1)
Sheriff	0.3***	0.4***	0.2***	0.3***
	(0.1)	(0.1)	(0.1)	(0.1)
Neighborhood	0.1	0.1*	−0.0	0.1
	(0.1)	(0.1)	(0.1)	(0.1)
Community	0.2***	0.1*	0.1	0.1*
	(0.1)	(0.1)	(0.1)	(0.1)
Peer	0.2***	0.1**	0.1	0.1**
	(0.1)	(0.1)	(0.1)	(0.1)
Duty	0.1*	0.1**	0.0	0.1
	(0.1)	(0.1)	(0.1)	(0.1)
AIC	24493.1	26068.9	21605.6	26093.5
BIC	24556.0	26131.7	21668.4	26156.3
Log Likelihood	-12238.6	-13026.4	-10794.8	-13038.7
Deviance	24477.1	26052.9	21589.6	26077.5
Num. obs.	19028	19028	19028	19028

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$. Holdout values in levels; remaining figures relative to this

Table A4: Logit Estimates Including Multiple Owners

	All Owners		Unary Owners	
	One Month	Three Months	One Month	Three Months
Lien	0.21***	0.20***	0.23***	0.22***
	(0.06)	(0.05)	(0.06)	(0.06)
Sheriff	0.15**	0.19***	0.16**	0.20***
	(0.06)	(0.05)	(0.06)	(0.06)
Neighborhood	−0.05	−0.03	−0.09	−0.05
	(0.06)	(0.05)	(0.06)	(0.06)
Community	−0.02	−0.06	0.00	−0.04
	(0.06)	(0.05)	(0.06)	(0.06)
Peer	0.01	−0.03	0.01	−0.02
	(0.06)	(0.05)	(0.06)	(0.06)
Duty	−0.06	−0.01	−0.06	−0.01
	(0.06)	(0.05)	(0.06)	(0.06)
AIC	25179.24	26349.91	21922.44	23174.00
BIC	25234.33	26405.00	21976.61	23228.16
Log Likelihood	-12582.62	-13167.95	-10954.22	-11580.00
Deviance	25165.24	26335.91	21908.44	23160.00
Num. obs.	19333	19333	16940	16940

$^{***}p < 0.001$, $^{**}p < 0.05$, $^{*}p < 0.1$

Bibliography

George A Akerlof. Procrastination and obedience. *The American Economic Review*, 81(2):1–19, 1991.

Michael G Allingham and Agnar Sandmo. Income tax evasion: A theoretical analysis. *Journal of Public Economics*, 1(3-4):323–338, 1972.

TM Andrews, MJ Leonard, CA Colgrove, and ST Kalinowski. Active learning not associated with student learning in a random sample of college biology courses. *CBE-Life Sciences Education*, 10(4):394–405, 2011.

Joshua Angrist and Guido Imbens. Identification and estimation of local average treatment effects, 1995.

Sarah F Anzia and Terry M Moe. Collective bargaining, transfer rights, and disadvantaged schools. *Educational Evaluation and Policy Analysis*, 36(1):83–111, 2014.

Dale Ballou and Michael Podgursky. Returns to seniority among public school teachers. *Journal of Human Resources*, pages 892–912, 2002.

Jeffrey L. Barnett and Phillip M. Vidal. State and local government finances summary: 2011. Technical report, Census of Governments, 2013.

Roland Benabou and Jean Tirole. Laws and norms. Technical Report 17579, NBER Working Paper No. 17579, 2011.

Samuel Berlinski and Matias Busso. Challenges in educational reform: An experiment on active learning in mathematics. IDB Publications (Working Papers) 88335, Inter-American Development Bank, March 2015. URL `https://ideas.repec.org/p/idb/brikps/88335.html`.

Timothy J Besley, Anders Jensen, and Torsten Persson. Norms, enforcement, and tax evasion. Technical report, CEPR Discussion Paper No. DP10372, 2015.

Saurabh Bhargava and Dayanand Manoli. Psychological frictions and the incomplete take-up of social benefits: Evidence from an irs field experiment. *The American Economic Review*, 105(11):3489–3529, 2015.

Barbara Biasi. Unions, salaries, and the market for teachers: Evidence from wisconsin. 2017. doi: http://dx.doi.org/10.2139/ssrn.2942134. URL `https://ssrn.com/abstract=2942134`.

Donald Boyd, Hamilton Lankford, Susanna Loeb, and James Wyckoff. Explaining the short careers of high-achieving teachers in schools with low-performing students. *The American Economic Review*, 95(2):166–171, 2005.

Donald J Boyd, Pamela L Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff. Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, 31(4):416–440, 2009.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

David Card, Stefano DellaVigna, and Ulrike Malmendier. The role of theory in field experiments. *The Journal of Economic Perspectives*, 25(3):39–62, 2011.

Lucio Castro and Carlos Scartascini. Tax compliance and enforcement in the pampas: Evidence from a field experiment. *Journal of Economic Behavior & Organization*, 116:65–82, 2015.

Raj Chetty, Adam Looney, and Kory Kroft. Salience and taxation: Theory and evidence. *The American Economic Review*, 99(4):1145–1177, 2009.

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

Raj Chetty, John N Friedman, and Jonah E Rockoff. Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *The American Economic Review*, 104(9):2633–2679, 2014b.

Michael Chirico, Robert P Inman, Charles Loeffler, John MacDonald, and Holger Sieg. An experimental evaluation of notification strategies to increase property tax

compliance: Free-riding in the city of brotherly love. *Tax Policy and the Economy*, 30(1):129–161, 2016.

Lora Cohen-Vogel, Li Feng, and LaTara Osborne-Lampkin. Seniority provisions in collective bargaining agreements and the teacher quality gap. *Educational Evaluation and Policy Analysis*, 35(3):324–343, 2013.

Yves Croissant. Estimation of multinomial logit models in r: The mlogit package. *R package version 0.2-2*, 2012. URL `http://cran.r-project.org/web/packages/mlogit/vignettes/mlogit.pdf`.

David B Dahl. xtable: Export tables to laTeX or HTML. *R package version 1.8.2*, pages 1–5, 2009.

Carl de Boor. *A Practical Guide to Splines*, volume 27. Springer-Verlag New York, 1978.

Louis Deslauriers, Ellen Schelew, and Carl Wieman. Improved learning in a large-enrollment physics class. *Science*, 332(6031):862–864, 2011.

Peter Dolton and Wilbert Van der Klaauw. The turnover of teachers: A competing risks explanation. *Review of Economics and Statistics*, 81(3):543–550, 1999.

Neil J Dorans. Correspondences between act and sat® i scores. *ETS Research Report Series*, 1999(1), 1999.

Matt Dowle and Arun Srinivasan. *data.table: Extension of 'data.frame'*, 2017. URL `http://r-datatable.com`. R package version 1.10.5.

Mimi Engel, Brian A Jacob, and F Chris Curran. New evidence on teacher labor supply. *American Educational Research Journal*, 51(1):36–72, 2014.

Thomas A Fischer, Wayne H Swanger, and Stacey Skoning. Supply & demand 2008: Data trends of education personnel in wisconsin public schools. Technical report, Wisconsin Educator Supply and Demand Project, 2009.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer, 2001.

Eleanor S Fulbeck. Teacher mobility and financial incentives: A descriptive analysis of denver's procomp. *Educational Evaluation and Policy Analysis*, 36(1):67–82, 2014.

Steven Glazerman, Ali Protik, Bing-Ru Teh, Julie Bruch, and Jeffrey Max. Transfer incentives for high-performing teachers: Final results from a multisite randomized experiment. *National Center for Education Evaluation and Regional Assistance*, 2013.

Dan Goldhaber, Betheny Gross, and Daniel Player. Are public schools really losing their best? assessing the career transitions of teachers and their implications for the quality of the teacher workforce. *National Center for Analysis of Longitudinal Data in Education Research*, 2007.

Dan Goldhaber, Lesley Lavery, and Roddy Theobald. Uneven playing field? assessing the teacher quality gap between advantaged and disadvantaged students. *Educational Researcher*, 44(5):293–307, 2015.

Peter J Green and Bernard W Silverman. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. CRC Press, 1993.

Michael Hallsworth. The use of field experiments to increase tax compliance. *Oxford Review of Economic Policy*, 30(4):658–679, 2014.

Michael Hallsworth, John List, Robert Metcalfe, and Ivo Vlaev. The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. National Bureau of Economic Research, 2014.

Eric A Hanushek and Steven G Rivkin. Constrained job matching: Does teacher job search harm disadvantaged urban schools? Technical report, National Bureau of Economic Research, 2010.

Eric A Hanushek, John F Kain, and Steven G Rivkin. Why public schools lose teachers. *Journal of Human Resources*, 39(2):326–354, 2004.

Douglas N Harris and Tim R Sass. Teacher training, teacher quality and student achievement. *Journal of Public Economics*, 95(7):798–812, 2011.

Andrew Haughwout, Robert Inman, Steven Craig, and Thomas Luce. Local revenue

hills: Evidence from four us cities. *Review of Economics and Statistics*, 86(2): 570–585, 2004.

Xuming He and Pin Ng. Cobs: Qualitatively constrained smoothing via linear programming. *Computational Statistics*, 14(3):315–338, 1999.

James J Heckman, Lance J Lochner, and Petra E Todd. Fifty years of mincer earnings regressions. Technical report, National Bureau of Economic Research, 2003.

Arne Henningsen and Ott Toomet. maxlik: A package for maximum likelihood estimation in R. *Computational Statistics*, 26(3):443–458, 2011. doi: 10.1007/s00180-010-0217-1. URL `http://dx.doi.org/10.1007/s00180-010-0217-1`.

Jessica B Heppen, Kirk Walters, Margaret Clements, Ann-Marie Faria, Cheryl Tobey, Nicholas Sorensen, and Katherine Culp. Access to algebra i: The effects of online mathematics for grade 8 students. *National Center for Education Evaluation and Regional Assistance*, 2011.

Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Brian A Jacob and Steven D Levitt. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, 118(3): 843–877, 2003.

Shanna Smith Jaggars and Thomas Bailey. Effectiveness of fully online courses for

college students: Response to a department of education meta-analysis. *Community College Research Center, Columbia University*, 2010.

Michael Keen and Joel Slemrod. Optimal tax administration. Technical report, NBER Working Paper No. 22408, 2016.

Judd B Kessler and C Yiwei Zhang. Behavioral economics and health. Technical report, Paper for Oxford Textbook of Public Health, 2014.

Henrik Jacobsen Kleven, Martin B Knudsen, Claus Thustrup Kreiner, Søren Pedersen, and Emmanuel Saez. Unwilling or unable to cheat? evidence from a tax audit experiment in denmark. *Econometrica*, 79(3):651–692, 2011.

William S Koski and Eileen L Horng. Facilitating the teacher quality gap? collective bargaining agreements, teacher hiring and transfer rules, and teacher assignment among schools in california. *Education*, 2(3):262–300, 2007.

Philip Leifeld. texreg: Conversion of statistical model output in R to LaTeX and HTML tables. *Journal of Statistical Software*, 55(8):1–24, 2013. URL `http://www.jstatsoft.org/v55/i08/`.

Steven D Levitt and John A List. What do laboratory experiments measuring social preferences reveal about the real world? *The Journal of Economic Perspectives*, 21(2):153–174, 2007.

Andrew Litten. The effects of public unions on compensation: Evidence

from wisconsin. 2016. URL https://drive.google.com/file/d/0BwL-PvOgW_Ordzk5QXloZGM1d1k/view.

Susanna Loeb and Marianne E Page. Examining the link between teacher wages and student outcomes: The importance of alternative labor market opportunities and non-pecuniary variation. *Review of Economics and Statistics*, 82(3):393–408, 2000.

Susanne Lohmann. The dynamics of informational cascades: The monday demonstrations in leipzig, east germany, 1989–91. *World Politics*, 47(01):42–101, 1994.

Betty Love, Angie Hodge, Neal Grandgenett, and Andrew W Swift. Student learning and perceptions in a flipped linear algebra course. *International Journal of Mathematical Education in Science and Technology*, 45(3):317–324, 2014.

Erzo F. P. Luttmer and Monica Singhal. Tax morale. *The Journal of Economic Perspectives*, 28(4):149–168, 2014.

Jean McGivney-Burelle and Fei Xue. Flipping calculus. *Primus*, 23(5):477–486, 2013.

Barbara Means, Yukie Toyama, Robert Murphy, Marianne Bakia, and Karla Jones. Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies. *US Department of Education*, 2009.

Terry M Moe. Bottom-up structure: Collective bargaining, transfer rights, and the plight of disadvantaged schools. *Education Working Paper Archive*, 2006.

Richard J Murnane and Randall J Olsen. The effects of salaries and opportunity costs on length of stay in teaching: Evidence from north carolina. *Journal of Human Resources*, pages 106–124, 1990.

Pin Ng and Martin Maechler. A fast and efficient implementation of qualitatively constrained quantile smoothing splines. *Statistical Modelling*, 7(4):315–328, 2007.

Ted O'Donoghue and Matthew Rabin. Doing it now or later. *American Economic Review*, 89(1):103–124, 1999.

Wisconsin Department of Public Instruction. School district name changes - since 1994. `dpi.wi.gov/sites/default/files/imce/sms/doc/rg_sdnamechanges.doc`, 2011. [Online; accessed 2017-05-13].

Wisconsin Department of Public Instruction. School staff: Salary, position & demographic reports. `https://dpi.wi.gov/cst/data-collections/staff/published-data`, 2017a. [Online; accessed 2017-02-22].

Wisconsin Department of Public Instruction. Data errata - data changes after dpi publication. `https://dpi.wi.gov/cst/data-collections/data-errata`, 2017b. [Online; accessed 2017-05-07].

Dina Pomeranz. No taxation without information: Deterrence and self-enforcement in the value added tax. *The American Economic Review*, 105(8):2539–2569, 2015.

Eric A Posner. Law and social norms: The case of tax compliance. *Virginia Law Review*, 86(8):1781–1819, 2000.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.

LA Reese and G Sands. No easy way out: Detroits financial and governance crises. In *Interrogating Urban Crises Conference, De Monfort University, Leicester, England, September*, 2013.

Meredith P Richards. The gerrymandering of school attendance zones and the segregation of public schools: A geospatial analysis. *American Educational Research Journal*, 51(6):1119–1157, 2014.

Steven G Rivkin, Eric A Hanushek, and John F Kain. Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458, 2005.

Jonah E Rockoff. The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, 94(2):247–252, 2004.

RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2017. URL `http://www.rstudio.com/`.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Carl Runge. Über empirische funktionen und die interpolation zwischen äquidistanten ordinaten. *Zeitschrift für Mathematik und Physik*, 46(224-243):20, 1901.

Jennifer Sable. Documentation to the nces common core of data local education agency universe survey: School year 2006-07 (nces 2009-301). *U.S. Department of Education*, 2009.

Benjamin Scafidi, David L Sjoquist, and Todd R Stinebrickner. Race, poverty, and teacher mobility. *Economics of Education Review*, 26(2):145–159, 2007.

Joel Slemrod. Cheating ourselves: The economics of tax evasion. *The Journal of Economic Perspectives*, 21(1):25–48, 2007.

Todd R Stinebrickner. An analysis of occupational change and departure from the labor force: Evidence of the reasons that teachers leave. *Journal of Human Resources*, pages 192–216, 2002.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 111–147, 1974.

Grace Wahba. *Spline Models for Observational Data*. SIAM, 1990.

Abraham Wald. The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics*, 11(3):284–300, 1940.

Andrew J Wayne, Kwang Suk Yoon, Pei Zhu, Stephanie Cronen, and Michael S Garet.

Experimenting with teacher professional development: Motives and methods. *Educational Researcher*, 37(8):469–479, 2008.

Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2016. URL `http://yihui.name/knitr/`. R package version 1.15.1.

Di Xu and Shanna Smith Jaggars. Online and hybrid course enrollment and performance in washington state community and technical colleges. *Community College Research Center, Columbia University*, 2011.

Di Xu and Shanna Smith Jaggars. Adaptability to online learning: Differences across types of students and academic subject areas. *Community College Research Center, Columbia University*, 2013.

Achim Zeileis. Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software*, 11(10):1–17, 2004. URL `http://www.jstatsoft.org/v11/i10/`.

Achim Zeileis. Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, 16(9):1–16, 2006. URL `http://www.jstatsoft.org/v16/i09/`.

Achim Zeileis and Torsten Hothorn. Diagnostic checking in regression relationships. *R News*, 2(3):7–10, 2002. URL `https://CRAN.R-project.org/doc/Rnews/`.

Stephen P Zeldes. Consumption and liquidity constraints: An empirical investigation. *Journal of Political Economy*, 97(2):305–346, 1989.