Essays on Teacher Labor Markets

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Kristine Lamm West

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Joseph Ritter and Paul Glewwe

October, 2012

Acknowledgements

Many thanks are due to my committee, Joe Ritter, Paul Glewwe, Elton Mykerezi and Aaron Sojourner, for their time, talents and excellent advising as well as Karine Moe and other professors at Macalester College for early and continued encouragement. I am indebted to Rachel Simmons for excellent research assistance and, to her and other former students and colleagues from Washburn High School for inspiration. I gratefully acknowledge funding from the graduate school, the National Council on Teacher Quality, the Minnesota Population Center and the Joseph M. Juran Center. Lastly, I would not have been able to complete this dissertation without the love and support of my family, most notably my parents, Foster and Didi, my husband, Jim.

Dedication

To my grandmother, Marjorie Lamm, who was adamant she wanted to enroll in Economics courses when they assumed she meant "Home Ec."

Abstract

This dissertation is comprised of three essays related to teacher labor markets. The first essay describes a theoretical model which incorporates an oft overlooked fact of educational production, namely the fact that teachers are asymmetrically well informed about what actions are best for their specific classes. The model shows that to take advantage of teachers' local knowledge, districts should offer contracts with output-based pay for performance coupled with decentralized decision making and support for teachers to help them set locally appropriate goals. I use data from Minnesota's Q-Comp program to empirically test the model. The data, however, do not confirm (or reject) the theory. The second essay investigates the impact of collective bargaining on teacher contracts using the 2003-04 and 2007-08 Schools and Staffing Survey (SASS) and data from a survey that I administered. Contracts negotiated via collective bargaining have greater returns to experience than do districts without collective bargaining. Unions do not appear to be a roadblock to basing compensation on student performance but they do oppose basing compensation on administrator review and basing tenure on student performance. The third essay turns to an analysis of average hourly wages. Using the American Time Use Survey (ATUS), I compare teachers' wages to demographically similar workers in other occupations. First I estimate that teachers work an average of 34.5 hours per week annually. Using the ATUS data, I conclude that high school teachers earn approximately 11% less than full time college educated workers in other occupations; but elementary, middle and special education teachers are not underpaid relative to full time college educated workers in other occupations.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

There are copious statistics cited in the academic and popular press that point to a failure to fully prepare our future workforce and to socio-economic disparities that pervade our educational system. We are faced with low levels of math and reading competency and achievement gaps between various demographic groups at all levels.[1] For instance, in 2009, only 33 and 39 percent of 4th graders performed at or above the "proficient" achievement level on the National Assessment of Education Progress (NAEP) in reading and mathematics respectively. Amongst black and Hispanic students less than 20 percent score "proficient" or above. Achievement levels do not improve (nor do achievement gaps narrow) as students move through our K-16 educational system. This is indicative of a deficit in human capital accumulation that has widespread implications and as such has attracted much attention from economists.

Recent and influential econometric studies have focused on the significant role that teachers play in student achievement (Rockoff, 2004; Nye et al., 2004; Rivkin et al., 2005; Aaronson et al., 2007). These studies use teacher fixed effects to estimate the test score improvement attributable to each teacher, which is referred to as the teacher's value added. This measure of a teacher's impact takes on many names in the literature - "teacher effectiveness" and "teacher quality" are the most common. Problems arise, however, when

[1] All statistics in this paragraph are from the National Center for Education Statistics.

policy discussions muddle these and similar terms. For instance, the No Child Left Behind Act talks about "highly qualified" teachers which refers to having met full licensure requirements rather than a teacher's value added. Even in academic papers, "teacher quality" can refer to multiple things at once, often encompassing observable and unobservable teacher characteristics. Further, using the term "teacher quality" infers that the traits of a good teacher are inherent while it may be better to conceptualize a teacher's value added as a combination of innate ability, learned skills and choices about effort. For these reasons, and because it is more in keeping with standard labor economic theory, I prefer "teacher productivity" and I will use this term throughout.

Regardless of the terminology, the consensus from value added research is that teachers are the single most important policy malleable factor in student achievement. For example, Nye et al. (2004) use data from the Project STAR experiment in Tennessee, and find that differences in teacher productivity are statistically significant and of policy relevant magnitudes. They find that the difference in achievement between having a 50th percentile teacher and a 90th percentile teacher is 0.33 standard deviations in reading and 0.46 standard deviations in math. By way of comparison, reducing class size from 25 to 15 (the initial goal of Project STAR) has only a 0.10 standard deviation effect. In addition, Nye et al. (2004) found that teacher effects are much larger in schools where students are of low socio-economic status, suggesting that focusing on teacher productivity may have an impact on differences in achievement that often are attributed to student demographics and help to close the achievement gap.

This research has quickly entered the public discourse. Bill Gates cites value added studies to motivate his foundation's efforts and President Obama often alludes to this research in his speeches on education. Perhaps the clearest indicator of teacher productivity's prominence in the world of public policy is the amount of money that has been allocated to the issue even in a time of deep budget cuts. The Race to the Top Fund offers in excess of $4 billion dollars in part to aid in efforts to "recruit, reward and retain quality teachers" (Race to the Top Program description). The competitive grant nature of this program has spurred stakeholders in every state to evaluate and invest in teacher productivity.

As a former teacher, I encounter this national focus on teachers with mixed emotions. On one hand there seems to be a tendency to blame teachers for the nation's educational

shortcomings. On the other hand, it is empowering to focus on the fact that highly productive teachers can have a significant impact on student achievement. I feel strongly that my research, grounded in both my training as a teacher and as an economist focused on labor and policy analysis, can make a valuable contribution to the national conversation about teacher productivity.

Discussions and policies about teacher productivity revolve around a number of important teacher labor supply decisions. Who becomes a teacher, how much and how hard they work, and whether they stay in the profession are all issues that require an understanding of teachers as labor market participants. There are fundamental questions that have not been adequately addressed in the research. Despite significant amounts of research and money devoted to improving teacher productivity, there remains a deficit in our knowledge about how teachers make labor supply decisions, how their choices compare to those of other workers, and how teacher contracts influence productivity. This deficit in knowledge has led to policies aimed at improving teacher productivity that are based on theory and empirical research borrowed from other sectors that do not take into consideration the unique features of educational production. Teacher labor markets and teacher contracts are unique in a number of ways, and my dissertation aims to fill some of the gaps in our knowledge. This research will be useful because it informs the larger policy discourse about teachers and because it addresses core issues such as labor market sorting, wage differentials, collective bargaining, time allocation and optimal contract design that are of interest to labor economists more broadly.

1.2 Context

The teacher labor market is an interesting labor economics case study for a number of reasons.[2] First, teaching is a large profession - there are over 3.7 million elementary and secondary teachers in the United States and more than 86% of those workers are employed in the public sector. Fully 10% of all college educated workers are teachers. There are roughly 14,000 public school districts spanning all types of communities. This provides both a large population of workers and considerable variation in work environments. Although

[2]Statistics are from the National Center for Education Statistics unless otherwise noted.

occupational licensure requirements and the institution of tenure (which is generally linked to job experience in a single district) present some barriers to movement, there are enough teaching opportunities to allow for considerable labor market mobility.

Secondly, teaching, especially elementary teaching, has historically been a largely female occupation (76% of all public school teachers are women and 15% of all female college graduates are teachers). Changing labor market opportunities for women therefore have a considerable impact on teacher labor markets. While the link between general scholastic aptitude and teacher productivity is not perfect, the two do appear to be correlated. Therefore it is concerning that, amongst college graduates, scholastic aptitude is inversely related to the choice of teaching as a profession (Podgursky et al., 2004; Goldhaber and Liu, 2002; Hanushek and Pace, 1995) and that the scholastic aptitude of teachers has declined over time (Corcoran et al., 2004). Many attribute this trend to the fact that college educated women now have higher wage opportunities in traditionally male occupations (Bacolod, 2007). Alternatively, Hoxby and Leigh (2004) argue that unionization and a resulting pay compression are to blame for the lack of high aptitude teachers.

Indeed, third and fourth unique aspects of teacher labor markets are the high degrees of both unionization and pay compression. Since the 1960s, teachers unions have become an increasingly important part of the educational landscape. In an era of declining union membership in most sectors, the share of teachers who are unionized has risen to 67%, up from just 22% in 1974 (Eberts, 2007). The National Education Association (NEA) and the American Federation of Teachers (AFT) are two of the most powerful labor organizations in the nation. One of the most notable changes that unionization brought to teaching was the single salary schedule. Initially intended to protect women and minority teachers from wage discrimination, the "steps and lanes" of teacher contracts define teachers' wages as a function of only two variables: education and experience. Unfortunately these two variables have been shown to be poor proxies for teacher productivity (Hanushek, 2003) and as a result many have called for reforms that take into consideration other measures. For example, in Washington D.C. a highly publicized reform effort both does away with traditional steps and lanes and tenure protections and replaces them with annual evaluations and pay linked to student achievement.

Lastly, there are important non-pecuniary aspects of teaching that potential and current educators consider when making labor supply decisions. Issues such as student behavior and parental support vary across schools creating very different work environments. Hanushek and Rivkin (2006) show that teacher transitions are more strongly related to student characteristics than to salary differentials. There are also time use benefits (often referred to in sociology as work-life benefits) to teaching. The most obvious benefit is the nine month academic year as the popular saying goes "the best three things about teaching are June, July and August!" Although they get summers off, anecdotal evidence suggests that teachers spend at least some of that time working to prepare for the coming school year and that during the school year they often take work home.

1.3 Essays

In this dissertation, I address three different, but inter related, topics on the economics of teacher labor markets. Each chapter is intended to be a stand alone essay. In the concluding section of this dissertation I draw connections between the three and discuss my plans for future research.

- Chapter 2: Complementarity of pay for performance and decentralized decision making

  Teachers are asymmetrically well informed about what actions are best for their specific classes: they have local knowledge about their own abilities and preferences, their students' abilities and preferences and the quality of the match between these. This essay describes a theoretical model that fits this environment. The model shows that to take advantage of teachers' local knowledge, districts should offer output-based pay for performance (P4P) coupled with decentralized decision making and support for teachers to help them set locally appropriate goals. I use data from Minnesota's Q-Comp program to empirically test the model. Q-Comp is an attractive setting to test the model because participating districts enacted a variety of pay and management reforms. Most notably, Q-Comp includes both pay for performance and support for decentralized decision making. I am unable, however, to confirm (or refute) the theory using this data. I attribute this to measurement error and

estimation issues.

- Chapter 3: Teachers' unions, compensation and tenure

  Teachers' unions have been targeted by a recent wave of legislation that limits collective bargaining rights. Unfortunately, behind the heated rhetoric that has accompanied these law changes, there is little formal research on the impact of unions on teacher compensation and tenure. This essay attempts to fill that void. First, data from the 2003-04 and 2007-08 Schools and Staffing Survey show that contracts negotiated via collective bargaining have greater returns to experience than do districts without collective bargaining. Districts that are unionized, either with or without legal collective bargaining protections, have higher returns to degrees and higher starting salaries than do districts without a union. Second, I argue that a commonly used, but very vague, definition of P4P obscures important differences in P4P contracts and the union response to such contracts. I survey human resource professionals from over 400 districts and find little evidence that unions oppose basing compensation on student performance. I find stronger evidence that unions oppose basing compensation on administrator review. Lastly, while almost all districts consider administrator review when granting tenure, unionized districts are slightly less likely to consider student performance or peer review when granting this important measure of job security. In sum, contracts negotiated via collective bargaining do appear to differ from those in non-union districts, but perhaps in some different ways than union opponents might predict.

- Chapter 4: Are teachers overpaid or overworked? New measures of market hours

  Researchers have good data on teachers' annual salaries but a hazy understanding of teachers' hours of work. This makes it difficult to calculate an accurate hourly wage and leads policy makers to default to anecdote rather than fact when debating teacher pay. Those who argue that teachers are overpaid point to a short contractual work day and year. Those who argue that teachers are overworked point to unpaid evenings, weekends and summers spent planning, grading and helping with extra curricular activities. Time diary data has the potential to settle this. Using data from the American Time Use Survey (ATUS), I find that teachers work an average of

34.5 hours per week annually (approximately 38.0 hours per week during the school year and 21.5 hours per week during the summer months). When hours per week are accurately accounted for, high school teachers earn approximately 11% less than full time college educated workers in other occupations; but elementary, middle and special education teachers are not underpaid relative to full time college educated workers in other occupations.

Chapter 2

Complementarity of Pay for Performance and Decentralized Decision Making.

2.1 Introduction

Although pay is the most high-profile aspect of negotiations between labor and management, personnel economics has long held that delegation of responsibility, monitoring, evaluation and training are all potential complements (or substitutes) for monetary rewards (Bloom and Van Reenen, 2011). Therefore, it is often misguided to negotiate changes to compensation schemes without simultaneously considering changes to management practices. In education there is a significant policy push to reform teacher compensation and move towards "pay-for-performance" (P4P). Unfortunately, these reforms are generally discussed without regard to other aspects of human resource management. This paper attempts to fill that void.

P4P contracts in education are diverse. Although economists may assume that all P4P contracts link teacher pay to student outcomes, the reality is much different. The vast majority of P4P contracts in education do not link compensation to student outcomes. Instead most link compensation to teacher actions such as participating in professional development and scoring highly on subjective evaluations. In short, anything other than

the traditional steps and lanes contract is apt to be termed P4P.[1] In what follows, I refer to both input and output based P4P. Output based P4P includes any contract that rewards teachers for student outcomes. Input based P4P refers to any contract that rewards teachers for their own actions including evaluation based P4P.[2]

Output based P4P contracts in education are particularly controversial. Both proponents and opponents have theory on their side. Those who argue in support of output based P4P point to basic principal-agent theory and lament the lack of incentives for teachers. If teachers' preferences are not closely correlated with the district's, linking teacher pay to student outcomes is necessary to extract optimal effort. Those who argue against output based P4P cite concerns about the potential for poorly constructed incentives to narrow the curriculum and erode cooperation among teachers. Linking teacher pay to student outcomes introduces dangerous incentives for cheating and other unproductive actions. Further, some argue that P4P is demeaning to teachers since they are committed to students' best interests regardless of the compensation scheme.

In this paper, I suggest using a theoretical model from Prendergast (2002) that describes the tradeoffs inherent in output based P4P for teachers. This model differs from the standard principal-agent model in how uncertainty and the production technology are characterized. To reflect this I use the term "technological uncertainty" to mean uncertainty about what particular inputs will produce. I develop intuition for the model as it relates to teaching and show that this way of thinking about uncertainty and the production technology has important implications for P4P in education. This approach unifies the arguments for and against output based P4P in education and brings into focus the positives and negatives of these contracts.

This theoretical model takes into account education's unique production technology. Specifically, teachers are asymmetrically informed about themselves and their classes and thus better placed than district officials to make choices about curriculum and pedagogy. Teachers know their own preferences and abilities, the preferences and abilities of their students and the quality of the match between these. The theoretical models that advocates

[1] Steps and lanes refers to the standard teacher contract in which compensation is determined by a set schedule where years of experience (i.e. steps) and years of education or educational degrees (i.e. lanes) are the only factors considered.

[2] There is no standard definition of output and input based P4P in education. Later I discuss the classification of different types of P4P with respect to Q-Comp in more detail.

and detractors of output based P4P in education most commonly cite do not incorporate the importance of teachers' local knowledge. This omission leads to conclusions that either over or understate the likelihood that output based P4P contracts will be successful in education. Using Prendergast's (2002) framework I am able to (1) underscore the assumptions needed for output based P4P to be successful in education and (2) suggest complementary management practices that will support output based P4P for teachers.

I show that to extract optimal effort, output based P4P must be coupled with decentralized decision making where the agent, in this case the teacher, has the authority to make decisions about which task to pursue. That is, output based P4P and decentralized (a.k.a. delegated) decision making are complements.[3] Districts that use output based P4P contracts should also decentralize decision making and provide support for teachers to use their local knowledge about what is best for their specific classes. I empirically investigate the complementarity of output based P4P and decentralized decision making using evidence from Minnesota's Quality Compensation program (Q-Comp). I exploit variation in Q-Comp at the district level in the timing of adoption, the design of P4P reforms, and the strength of complementary management practices to test the predictions of the model and provide new evidence on optimal contracting. Unfortunately, the results are inconclusive likely, because of measurement error and estimation issues which are discussed.

The chapter proceeds as follows. Section 2.2 summarizes the theoretical model and builds intuition for how it relates to education. Section 2.3 provides evidence that education fits the key assumptions of the model. Section 2.4 discusses the unique features of Q-Comp and the data available. Section 2.5 outlines the empirical strategy and section 2.6 presents the results. Section 2.7 concludes with applications to current educational policy debates and suggestions for future research.

[3] I use the terms decentralized decision making and delegated decision making interchangeably. Seminal authors in this area include Radner (1993) who discusses when "decentralized information processing" is optimal and Aghion and Tirole (1997) who discuss when "delegated formal authority" is optimal. In a broader sense, my paper describes an incomplete contracting approach (Grossman and Hart, 1986) where the contract between the principal and the agent is not over the specific task to be performed but rather over who will have the right to decide what task is rewarded and how the agent will be held accountable.

2.2 Theory

2.2.1 Intuition

Prendergast (2002) presents a principal-agent model where the production technology is such that the principal does not know which input is most productive and thus must delegate the choice of input to the agent. The agent chooses one input out of a set of possible alternatives. From the principal's point of view there is a lot of uncertainty about which input is best. The agent, however, is asymmetrically well informed. I refer to this type of uncertainty as "technological uncertainty" to distinguish it from the standard treatment of uncertainty that focuses on the agent's risk aversion.[4] This model fits the educational production technology well. Teachers choose one lesson out of a large set of alternatives. The school district does not know which lesson is best. Teachers are asymmetrically well informed because they know more than the district does about their abilities and preferences, the abilities and preferences of the students in their classes and the quality of the match between these variables and all the possible lessons they can teach on a given day. This is the teachers's "local knowledge."

Prendergast's model builds on a well known extension of the classic principal-agent problem that describes an environment where the agent's job is characterized by many tasks (Holmstrom and Milgrom, 1991). In this type of job environment it is difficult to construct a contract that provides incentives for the agent to exert effort on the various tasks in the mix that the principal desires. This "multi-tasking" model is often used as an argument against paying teachers for student outcomes.[5] The argument is that since teaching is a multi-dimensional job, rewarding teachers for observable outputs will

[4]Prendergast's treatment of uncertainty is different than how principal-agent models usually treat uncertainty. Standard principal-agent models predict a negative relationship between uncertainty and incentives. This is because a risk-averse agent requires higher expected wages to compensate for uncertainty. Although the theory predicts a negative relationship, empirically, the opposite is often observed. P4P actually seems to be *more* prevalent in jobs where outcomes are very uncertain (such as CEOs and franchisees). Prendergast argues that the standard models fail to consider the effect of uncertainty on other aspects of job design. Specifically, in very uncertain environments, the choice of task is delegated to the agent so the agent can make use of local knowledge and P4P is used to hold the agent accountable for that choice. By characterizing uncertainty as overlap in the distribution of potential outcomes for discrete choices about tasks, he is able to show a positive relationship between uncertainty and incentives. In what follows, I provide an illustration of this tailored to educational production.

[5]In fact, Holmstrom and Milgrom (1991) use teachers and teaching to the test (or even cheating) as the motivating example in their seminal paper.

cause them to ignore equally important, but harder to measure, outputs. Simply put, paying for observable output will provide incentive for teachers to teach to the test. Using Prendergast's model I show, however, that the fact that teaching is a multi-dimensional job can also be a reason to prefer output based pay. This result stems from the fact that teachers have important information and paying for outputs provides an incentive for teachers to utilize this information effectively. In the end, we must weigh the costs of teaching to the test against the benefits of the teachers' local knowledge.

Consider the following illustration to build on this intuition. There are two lessons: j and k. The teacher knows the efficacy of each possible lesson for her specific class. However, from the district's point of view the efficacy of each possible lesson is random. The district only knows the average efficacy of each lesson across all classes, μ_j and μ_k, and the variance, σ^2.[6] The variance is a measure of how much technological uncertainty there is. That is, for a given vector of means, σ^2 measures how sure the district is about which lesson is best. A high σ^2 means there is a lot of technological uncertainty – the district is rather unsure about which lesson is best and the teacher's local knowledge is very important. In Figure 2.1 panel I, I assume that the efficacy of each lesson is normally distributed and $\sigma^2 = 1$. The district knows that, on average, lesson k is the better choice, however, the overlap in the distributions means that it is possible that the actual realizations, marked by bold "j" and "k" on the illustration, will make lesson j the better choice.[7]

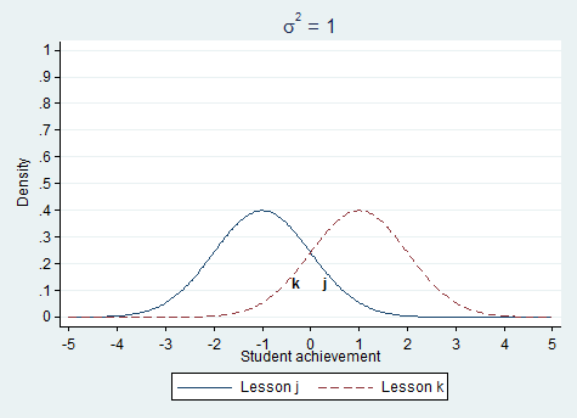
In Figure 2.1 panel II, there is less technological uncertainty, $\sigma^2 = 0.5$. In this case, the district is fairly sure that lesson k is the better choice. It is unlikely that the teacher's local knowledge will change this conclusion. In Figure 2.1 panel III, there is a lot of technological uncertainty, $\sigma^2 = 2$. The district is very unsure and the teacher's local knowledge is very valuable. There is a high likelihood that, although lesson k is better on average, lesson j may turn out to be a better fit for the specific teacher and class.

[6] I could also use σ_k^2 and σ_j^2 to denote different variances for each lesson. I assume, however, the variance is constant across both lessons so I drop the subscript for simplicity.

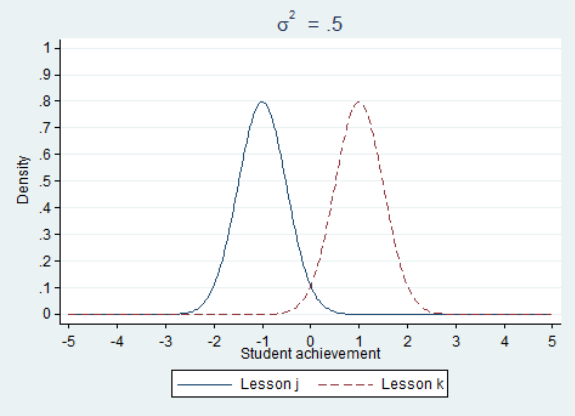
[7] This is an extreme version of a more general case where the teacher is also uncertain about the efficacy of each possible lesson but the teacher's beliefs have a tighter distribution than the district's beliefs.

Figure 2.1: Technological Uncertainty Example

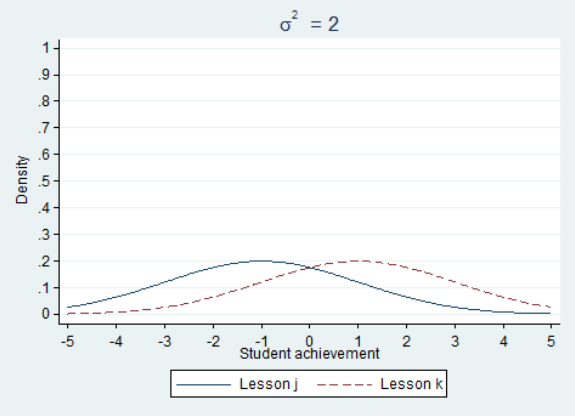
I: Moderate Uncertainty



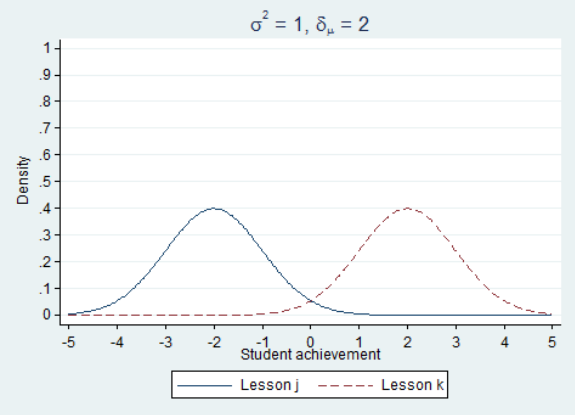
II: Less uncertainty (variance)



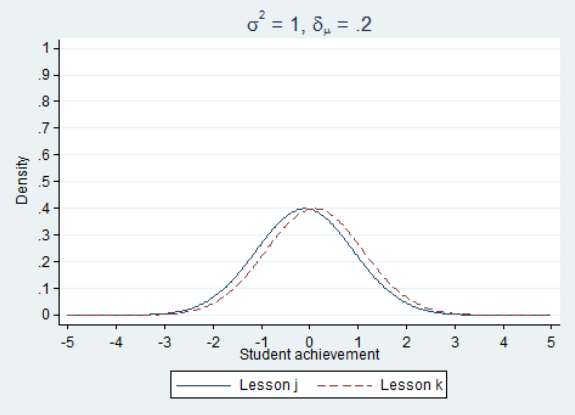
III: More uncertainty (variance)



IV: Less uncertainty (means)



V: More uncertainty (means)



Another way to characterize technological uncertainty is through the difference in the means. In this case, hold σ^2 constant and allow the means to be relatively far apart, Figure 2.1 panel IV, or relatively close, Figure 2.1 panel V.

The key observation is that when there is a lot of technological uncertainty, there is significant overlap in the distributions, and the district needs the teacher to act upon her asymmetrically held information. The main result of the model is that if there is little technological uncertainty (Figure 2.1 panels II and IV) the district should direct the teacher to teach lesson k. If there is a lot of technological uncertainty (Figure 2.1 panels III and V) the district should delegate the choice lesson to the teacher.[8]

2.2.2 Optimal Contract: The Role of Uncertainty

More formally, consider a district (the principal) who hires a teacher (the agent) to exert effort on one of N possible lessons, $i = 1...N$. The teacher exerts effort, e_i, at cost $C(e_i)$. This effort is observable to the district, i.e. via formal evaluations. Student achievement, the output, y_i, depends on effort and $lessonspecific$ unobservable teacher/student/class idiosyncrasies, ρ_i.

$$y_i = \rho_i + e_i \tag{2.1}$$

From the point of view of the district, ρ_i are random. The district knows only the distributions of ρ_i. As above, assume that all ρ_i have the same variance, σ^2, but differ in their means, μ_i. This models the fact that some lessons are better suited than others for a specific teacher and a specific class. The district has only a probabilistic understanding of which lesson is best and it would be costly to obtain more exact information. The teacher, on the other hand, knows (or learns before the district does) the realized value, $\widehat{\rho_i}$, for all i.

The teacher has preferences over lessons that are not perfectly correlated with the efficacy of the lesson. Thus there is a principal-agent dilemma and the district seeks to craft a contract that provides an incentive for the teacher to chose the best lesson rather than her preferred lesson. Assuming that the teacher does not necessarily prefer the lesson that

[8] One can think of this akin to power calculations used in hypothesis testing. If there is a high likelihood that the district will incorrectly accept the hypothesis that lesson k is better than lesson j, then they prefer to delegate the choice of lesson to the teacher.

is best for the class represents the fact that teachers, like most workers, have preferences that are not always perfectly (or even positively) correlated with the firm's. This problem motivates the entire principal-agent literature. Teachers have preferences over lessons for many reasons other than effectiveness. They may prefer lessons that are easy to prepare such as those they have already prepared in previous years or for other classes. They may prefer lessons that are easy to assess or ones that are easy to manage. They may also prefer lessons that they find personally interesting. The district does not know which lesson the teacher prefers and therefore believes that the distribution over the preferred lesson is uniform.

A contract between the teacher and the district is described by two features. First, there is the choice of whether to base the wage on evaluations which measure input, e_i, or student test scores which measure output, y_i. Second, there is the choice of whether the district will direct the choice of lesson or delegate the decision to the teacher.

The district can monitor observable effort at cost m_e or it can monitor observable output at cost m_y.[9] These can be direct costs such as an administrator's time to conduct an evaluation. They can also be indirect costs such as lost productivity due to misallocated effort. Paying based on student test scores may be expensive because teachers spend time teaching to the test rather than engaging students in meaningful lessons. I assume that $m_e < m_y$ to represent the fact that teaching to the test is potentially quite costly. Another way to think of this is that m_y is the cost of creating or administering a perfect test, i.e. one that is able to capture true learning. This test would be very costly to create and/or administer.[10]

If the teacher is indifferent about which lesson to teach, then there is no agency dilemma. The district could hire the teacher and extract her knowledge of which lesson is best at no cost (i.e. just ask) and, since effort is observable, either payment based on input or output will have identical results. If payment based on inputs is cheaper, as I have assumed, the district will offer an input based contract. If the teacher has preferences over which lesson to teach, the district needs a mechanism to motivate the teacher to teach the most effective lesson rather than her preferred one. We can draw on principal-agent theory to describe a contract that achieves this. The district will offer a contract that maximizes

[9] For simplicity, I assume that the district can monitor m_e or m_y, not both.

[10] In this case we could write $m_y(a)$ where a is accuracy and $m_y^{'} > 0$, that is, cost is increasing in accuracy.

surplus given the assumption that the teacher will respond rationally. Again, the district has two decisions to make when designing a contract. It must decide whether to pay based on inputs or output and whether to direct the choice of lesson or delegate that decision to the teacher. The district's and the teacher's maximization problems are outlined in the appendix along with a detailed comparison of the four different contract types. Here I present the conclusion and intuition.

For sake of argument, let the realization of $\widehat{\rho_j} > \widehat{\rho_k}$, i.e. the district would have chosen the wrong lesson. The main conclusion from the comparison of contract options is that a district will prefer a contract that pays for output and delegates decision making if

$$\widehat{\rho_j} - \widehat{\rho_k} > m_y - m_e \tag{2.2}$$

The intuition behind this result is that, if the benefit from using output based pay to extract the teacher's local knowledge exceeds the cost of using output based pay, i.e. costs such as lost productivity from teaching to the test, then output based P4P is a good idea. The model makes clear that there is a trade-off inherent in output based P4P contracts. Supporters of output based P4P argue that the benefits outweigh the costs, $\widehat{\rho_j} - \widehat{\rho_k} > m_y - m_e$, while those who oppose output based P4P argue that this inequality does not hold.

The model shows that the optimal contract depends on the relative costs of monitoring inputs and outputs and how asymmetrically well-informed the teacher is. The marginal benefit of delegating the choice of lesson (or the marginal cost of directing the choice of lesson) is the distance of $\widehat{\rho_j}$ from $\widehat{\rho_k}$ and this depends the level of technological uncertainty in the environment. This is shown by the amount of overlap between the distributions, either due to a large σ^2 or a small difference in the means. As illustrated above, the overlap in the distributions measures how likely the district is to choose the wrong lesson. If there is a lot of technological uncertainty, the expected benefit of choosing the best lesson is large. If there is little technological uncertainty, the expected benefit of the right lesson is small.

I assume $m_y > m_e$ in education because output based pay may lead to unproductive actions such as teaching to the test so the right hand side of equation (2.2) is non zero. This cost must be weighed against the benefit of choosing the right lesson. If there is

sufficient technological uncertainty, the benefit of choosing the right lesson outweighs the cost of teaching to the test and the district should pay based on outputs and delegate the choice of lesson. In other words, output based P4P and delegated decision making are complements. Coupling output based P4P with delegated decision making provides both the incentive and the ability for teachers to chose the best lesson for their specific classes. Delegated decision making alone allows teachers to choose the lesson they prefer rather than the one that maximizes output. On the other hand, output based P4P alone tells teachers to "do better" but does not give them the ability to choose the best lesson.

2.3 P4P in Education

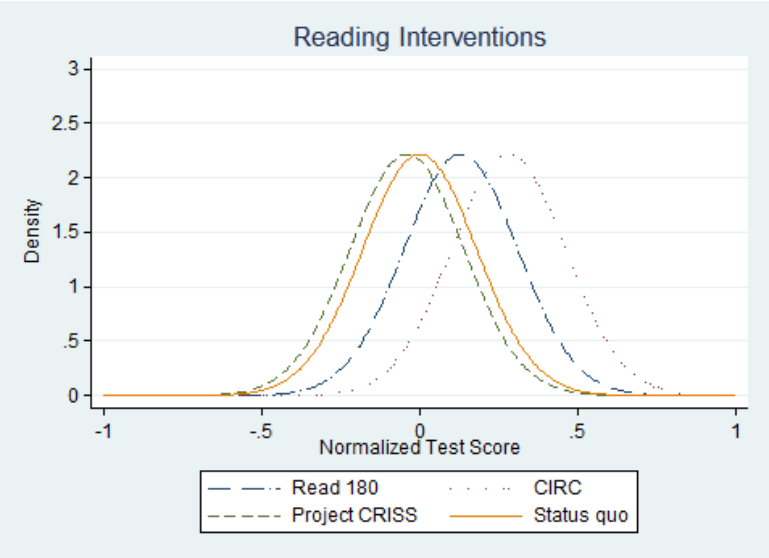
The theoretical model clearly shows that in multidimensional work environments there is a tradeoff between the benefit of motivating the agent to reveal local knowledge and the cost of overemphasizing the rewarded metric. The model predicts that output based P4P contracts that delegate decision making responsibilities will be optimal under certain conditions. Therefore, I now ask whether the model is appropriate for education. There are three key questions: (1) Is there significant technological uncertainty in education? (2) Are teachers asymmetrically informed about what is best? and (3) What is the nature of monitoring in education? Once I address these questions, I move to a discussion of the status quo in teacher contacts. I conclude that the status quo is not optimal and consider alternatives.

2.3.1 Is there Significant Technological Uncertainty in Education?

If there were a single "best-practice" curriculum and/or pedagogical practice that worked equally well for all students, there would be no need to delegate decision making responsibility to teachers. The district could assign a lesson to the teacher and pay based on formal evaluations that measure teacher effort. There is, however, considerable uncertainty about what is best in education.

The U.S. Department of Education's Institute of Education Sciences maintains a database of research on educational interventions called the What Works Clearinghouse. I use the

Figure 2.2: Uncertainty in Education



database to identify research-tested curriculum and pedagogy. For the purpose of illustration I focus on elementary reading, but the results of the analysis are the same for other subjects and/or grades. The three interventions for elementary reading with the most research backing are Read 180, a computer program designed to track and adapt to each student's progress; Project CRISS, a professional development program for teachers based on cognitive psychology and brain research; and CIRC, a curriculum based on daily lessons that provide students opportunities to practice comprehension and reading skills in pairs and small groups.[11]

Figure 2.2 illustrates the effectiveness of each of these compared to the status quo. I use the effect size from the largest study of each intervention. Effect sizes normalize the outcome to mean zero and standard deviation one to ease comparison across studies so

[11]The methodology for choosing these interventions was that they met the following criteria on the What Works Clearinghouse search engine: Reading/Writing, grades 3-8, general education, potentially positive effects, extent of evidence: medium to large, delivery: whole class, curriculum.

here the uncertainty is shown as differences in means.[12] In education, an effect size of 0.3 is considered reasonably large. The effect sizes for Read 180, Project CRISS and CIRC are 0.28, 0.06 and 0.49 respectively. The effect size of these interventions are, however, very small relative to the dispersion, i.e. there is a lot of overlap in the distributions. Figure VI makes clear that education is characterized by considerable technological uncertainty and thus is exactly the setting where there are potentially large gains from making good use of the agent's local knowledge.

2.3.2 Are Teachers Asymmetrically Informed About What Is Best?

A necessary condition of the model is that the agent have local knowledge. That is, she must be asymmetrically informed about the efficacy of different inputs. It is not enough for there to be considerable uncertainty, teachers must have superior information about which lesson is best. Consider the three interventions in Figure VI, CIRC is the best choice in expectation, but it appears likely that one of the other choices may be better for a particular class.[13] Do teachers know whether Read 180, Project CRISS, CIRC or the status quo is best for their particular talents and their particular students?

Another way to ask this question is: How private is the teacher's information? The teacher may be asymmetrically informed, but if the district can learn what the teacher knows relatively cheaply, then it can solve the agency dilemma directly. One could imagine the district hiring a second agent who's job it is to gather the local information and report back to the district. In this case the district would learn the realizations of $\widehat{\rho}_i$ rather than simply the distributions. Recall, however, that the teacher's local information is a complex set of variables. It includes the abilities and preferences of the students, the abilities and preferences of the teacher and the quality of the match between the two. The realizations

[12] To be more exact, because effect sizes force the standard deviation to be one, I cannot say whether the underlying uncertainty is due to a large standard deviation or a small difference in means. The methodology of effect size normalizes the standard deviation and forces all of the difference into the mean. Here I adjust the standard deviation of the intervention to account for the fact that it is a classroom, i.e. teacher, level choice rather than a student level choice. Specifically, I divide the standard deviation by the square root of 30 (an average class size).

[13] A necessary assumption is that class-specific efficacies of the interventions are not correlated. That is, the realizations of $\widehat{\rho}_i$ are independent draws. If this assumption is violated a class that is better suited than the average class for Read 180 is also better suited for Project CRISS and CIRC, then the characterization of technological uncertainty as random draws is problematic. This is an area for future investigation.

of $\widehat{\rho_i}$ depend on all of these factors and it is likely prohibitively expensive to acquire local knowledge comparable to the teacher's. Teachers may require trial and error to determine what is best for their students. The critical question is whether teachers can determine what is best more easily than district administrators can.

I argue that teachers have the *potential* to be asymmetrically well informed and that complementary management practices can help them actualize this potential. In the case of Minnesota's Q-Comp, many districts implemented a complementary decentralized professional development process intended to aid teachers in figuring out what is best for their students. I discuss this at length in Section 4.

2.3.3 What is the Nature of Monitoring in Education?

I have assumed that it is more costly to measure output than it is to measure input. I now consider monitoring in education in more detail to probe whether this is an appropriate assumption.

The measure of output closest to y_i is arguably a teacher's value added (Hanushek and Rivkin, 2006). Therefore, the costs of monitoring output include developing and administering student achievement tests and using these to calculate meaningful teacher-level value added scores. Districts routinely administer standardized tests and the Elementary and Secondary Education Act (a.k.a. No Child Left Behind) requires annual testing for all grades. Districts, therefore, have the data necessary to measure teacher output using student-level data, however, there are at least two concerns: (1) standardized tests are imperfect measures of student learning and (2) teacher-level value added scores are imperfect measures of teacher quality (Lockwood et al., 2007; Rothstein, 2010).

Calculating meaningful value added scores would require districts to administer an accurate test *and* employ accurate statistical models that isolate the effect of the teacher, both of which are currently unavailable. Therefore, any contract that uses the currently available tests and models may motivate inefficient teacher actions. Research has shown that poorly constructed incentives lead to inefficient, unproductive hidden teacher actions in the form of coaching (Jacob, 2005), socially wastefully gaming (Figlio and Winicki, 2005) or even cheating (Jacob and Levitt, 2003). Each of these are a variant of the idea that "you get what you pay for," that is incentives for test scores will get you higher test scores,

but not necessarily real student learning.

Even if accurate tests and models were developed, many districts do not have the necessary data infrastructure nor statistical expertise needed to implement payment based on them. Additionally, value added scores are a controversial measure of teacher productivity that may be opposed by powerful interests such as teachers' unions (West and Mykerezi, 2011).[14] In sum, clearly there are nontrivial costs in terms of physical, human and political capital needed to move to output based contracts in education.

The most obvious costs of monitoring teacher inputs is the time needed to evaluate teacher effort. Currently most districts only very crudely monitor and reward effort and this requires minimal time. Once teachers earn tenure, they can be fired only if their attendance or behavior is egregious. Attendance and minimally adequate behavior are measures of input, but they set a low bar. Increasingly, districts are attempting to raise the bar by conducting multiple formal evaluations each year to monitor and reward teacher effort. This, however, is more time consuming and thus more costly. When pay is contingent on evaluations I term this input based P4P. Input based P4P is more attractive politically than output based P4P for a number of reasons: (1) evaluations can be used in non-tested subjects, (2) research has shown that principals are able to distinguish effective from ineffective teachers (Jacob and Lefgren, 2008; Rockoff et al., 2011; Tyler et al., 2010) and, (3) in theory, they can mitigate the aforementioned problems of teaching to the test (Baker et al., 1994).

In sum, the assumption that $m_e < m_y$ largely reflects that fact that measuring output is costly because it may lead to unproductive hidden teacher actions since we do not have a test that accurately measures true learning nor do we have reliable value added models that isolate teacher productivity. Even if we had accurate tests and models, there is a significant investment in physical, human and political capital needed for output based measurement to be feasible. Measuring inputs on the other hand is already crudely done and there are politically attractive enhancements to this management practice which, although costly in terms of evaluator time, may mitigate lost productivity due to teaching to the test.[15]

[14] One only need to witness the recent events in New York City to realize that unions have reason to be concerned that value-added data may be used in unintended ways.

[15] It is worth noting that evaluations are not without drawbacks. For instance they are easily corruptible if teachers and evaluators collude. Neal (2011) speculates that the failure of P4P programs in England (Atkinson et al., 2004) and Portugal (Martins, 2009) may have been due to the fact that they were largely

2.3.4 What is the Status Quo in Education Contracts?

Are teacher contracts currently based on inputs or outputs? Do they delegate or direct the choice of lesson? To answer these questions I turn to the National Center on Education Statistic's Schools and Staffing Survey (SASS). I conclude that the modal contract in education pays for inputs and delegates decision making.

Teachers are traditionally paid based on rigid steps and lanes contracts that reward only academic degrees and years of service; clearly teacher inputs. According to the SASS, in 2007-08 fewer than 10% of teachers worked in districts that offer "pay for excellence" (this is the closest question the SASS has to question about output based pay).

Teachers also have a good degree of autonomy. The SASS asks teachers how much control they have over curriculum and pedagogy. In 2007-08 over 65% report that they have a moderate or great deal of control over "selecting content, topics, and skills to be taught" and over 95% report that they have a moderate or great deal of control over "selecting teaching techniques."[16]

Is it the case that pay for excellence (i.e. output based pay) is correlated with the degree of control that teachers have over content and technique (i.e. delegated decision making)? That is, is there evidence in the SASS that schools pair output based pay and delegated decision making as the theoretical model predicts they should? There is not. Teachers who work in districts that "pay for excellence" are no more likely to report that they have control over decisions about curriculum and pedagogy than teachers in districts that do not "pay for excellence."[17]

2.3.5 Is P4P Optimal for Education?

To summarize: education is characterized by considerable uncertainty; teachers have at least the potential to be asymmetrically well informed; monitoring output is relatively expensive; decision making responsibility is generally delegated to teachers; and teachers

based on subjective evaluations done by local staff. Such plans may not improve student achievement because evaluators lack incentives to assess teachers accurately. In this case, lost productivity is due to collusion rather than unproductive hidden actions.

[16] Author's calculations using the public use version of the 2007-08 SASS.

[17] In districts that "pay for excellence," 65% and 95% of teachers report that they have a moderate or great deal of control over content and techniques, respectively. The corresponding numbers for teachers in districts that do not "pay for excellence" are 62% and 94% (author's calculations).

are usually paid for inputs. I can now ask whether the theoretical model predicts that output based P4P is optimal for education and discuss possible alternatives to the standard contract.

First note that the model predicts that delegating decision making responsibility and paying for inputs, the status quo in education, is optimal only if the teacher's preferences are perfectly correlated with the district's (or the teacher has no preferences other than monetary rewards). The model predicts that because teachers have preferences over lessons for reasons other than efficacy, if the district is going to pay based on inputs, it should direct the choice of lesson. Thus, one potential solution to the standard contract in education is to continue to pay based on inputs, but take decision making power away from teachers. This type of reform is popular with some. States such as Texas and California have adopted common curriculum and popular charter schools like KIPP give teachers little to no choice over curriculum and pedagogy.

The model predicts, however, that paying for inputs and directing the choice of lesson is dominated by paying for outputs and delegating the choice of lesson if there is considerable technological uncertainty. Given that most teachers already have control over the curriculum and pedagogy, adding output based P4P should work well in education. A recent review of the literature finds that output based P4P in education has shown some promise (Neal, 2011). It is not the case, however, that output based P4P in education has been universally successful. Historically, when districts try output based P4P, they generally abandon it after a few years (Murnane and Cohen, 1986). There is also reason to doubt the efficacy of output based P4P in education because two recent large scale randomized trials have found null effects. In Nashville, Tennessee teachers could earn up to $15,000 for gains in student achievement. After two years, student test scores for those in the treatment group were no better than those in the control group (Springer et al., 2010). In New York City, schools could earn bonuses of up to $3,000 per teacher based on a composite measure that included student achievement as well as data on attendance and discipline. After two years, student test scores, attendance and graduation rates in the treatment schools were no better than those in the control group. In fact, in New York City, the P4P may have actually *decreased* student achievement, especially in larger schools (Fryer, 2011).

Why has output based P4P not been universally successful in education? The theoretical model, along with careful consideration of the model's assumptions, provides a possible answer to this question. I have only been able to state that teachers have *the potential* to be asymmetrically well informed. What if teachers are asymmetrically informed but not asymmetrically *well* informed? In other words, teachers certainly have local knowledge about themselves and their students. They may not, however, know what lesson is the best fit. In the the following section I discuss complementary management reforms that may be needed to ensure that teachers can make good use of their local knowledge.

Interestingly, if this is true, the model predicts that teachers may choose the wrong lesson even if they have preferences that are perfectly correlated with the district. In this case the teacher's preferred lesson can be seen as the lesson the teacher thinks will work. Teachers and the district can have fully correlated preferences, teachers can exert the optimal level of effort but it will be the wrong effort because teachers are doing what they think is best not what is actually best. Thus, whether or not teachers have preferences over lessons for reasons other than efficacy, schools that implement output based P4P will also need to implement complementary reforms that help teachers make good use of their local knowledge. Output based P4P without these complementary reforms will not work. Shortly, I discuss an example in Minnesota that shows such complementary reforms do exist in current education policy initiatives.

2.3.6 Is Support for Decentralized Decision Making the Missing Piece?

If it is true that that teachers do not have a clear idea of which lesson is best, a district that implements output based P4P alone is simply telling teachers to "do better" without giving them the resources to figure out what to do. Given one of the most basic tenets in economics - that people respond to incentives - is it any surprise to see evidence that output based P4P in education leads to coaching, gaming, teaching to the test or even cheating? A teacher who is motivated by output based P4P will change her behavior to earn the reward. One option is for the teacher to devote time to unproductive hidden actions.

Another option is that teachers devote time to figuring out what the best lesson is for her class. Interestingly, in the "failed" Nashville experiment, researchers found that

teachers in the treatment group were significantly more likely to seek out opportunities for collaboration. Teachers in the treatment group reported that they collaborated more on virtually every measured dimension (Springer et al., 2010). I view this as evidence that teachers were motivated by the rewards and their actions reveal useful information about how to support teachers and help them use their local knowledge effectively. Output based P4P motivated teachers in Nashville to seek out collaboration. This reveals that they did not immediately know which lesson was best, but they decided that the best way to figure it out was to collaborate with their colleagues.

In Minnesota's Q-Comp program, output based P4P is coupled with a decentralized professional development process that emphasizes collaboration and supports teachers in setting and working towards individual or small team-level goals. I test whether districts that couple output based P4P with support for collaboration achieve gains in student reading achievement. If so this would be consistent with the theoretical model and support the assertion that output based P4P and management practices that provide support for teachers to collaborate and make good use of their local knowledge appear to be complementary reforms that should be implemented together.

2.4 Minnesota's Q-Comp

In 2005, the state of Minnesota implemented the Quality Compensation program (Q-Comp) as the signature education reform of Governor Tim Pawlenty. Since then, dozens of districts have participated with over one million student-years taught. In order to participate a district must apply to the state. The Minnesota Department of Education (MDE) set general guidelines and districts can propose specific programs. If the proposal is approved by the MDE and the local teachers' union, the state authorizes up to $260 per student per year in additional funding for the district.[18]

Q-Comp provides an excellent opportunity to learn about P4P in education for a number of reasons. Most importantly for this study, Q-Comp is not simply a P4P reform. The MDE requires that districts couple P4P with management reforms. This provides a tremendous opportunity to learn about the complementarity of P4P and management

[18]The state provides general education aid of approximately $6,100 per student per year so Q-Comp adds up to 4% to the average district's baseline funding per student.

practices. Q-Comp plans contain five components: (1) career ladders/advancement options; (2) job-embedded professional development; (3) teacher evaluation/observation; (4) P4P; and (5) an alternative salary schedule. I focus on the P4P, job-embedded professional development and evaluations components. More detail follows on each.

The P4P component is divided into bonuses for teacher-level or small team-level goals, school or district-level goals and formal evaluations. Districts vary in the amount of money at stake for each of these. The modal district offers up to $2,000 per teacher per year. On average, about $1,000 is tied to evaluations, $800 is tied to teacher or small team-level goals and $200 is tied to school or district-level goals, however there is considerable variation. Histograms and summary statistics are provided in Table 2.1.[19] Payouts for each category are generally binary, that is a teacher either earns a bonus or does not. No district has a linear payout scheme and only a few have an option to earn a partial bonus.

Only one district ties rewards to value-added measures. Instead of using value-added scores, the state encourages districts to link the teacher or small team-level P4P bonuses to the job-embedded professional development (JEPD) component of Q-Comp. JEPD is a decentralized approach to professional development where the district sets broad goals and then asks small teams of teachers to work together in pursuit of these goals. Teams of teachers meet regularly to discuss and analyze curriculum, pedagogy and student outcomes. These teams of teachers support each other in working towards a measurable, student-centered goal. JEPD is clearly designed to help teachers figure out and implement what works for their specific classes. I use JEPD as a measure of support for decentralized decision making.

Teachers earn bonuses for working towards and achieving their goal. Because districts consider both the teacher's actions (i.e. participating in the professional development) and student outcomes (i.e. achieving the measurable, student-centered goal), there are elements in the P4P component of Q-Comp of both input and output based P4P. Despite this, I use the bonuses for teacher or small team goals as a measure of output based P4P.

The state encourages districts to link the evaluation bonuses to formal observations conducted by the school principal, district mentors or peer evaluators. The state guides districts to use the Charlotte Danielson evaluation framework (Danielson and McGreal

[19] See Sojourner et al. (2012) for more detail about the P4P component of Q-Comp.

2000) which is well regarded in the education literature. Trained evaluators conduct at least three observations per teacher per year. Although the evaluations are subjective, they rely heavily on a rubric and the state stresses the importance of inter-rater reliability, thus they are very formal "subjective" evaluations. Teachers earn bonuses for participating in the evaluation process and for scoring highly on the observation rubric. Although they likely include some output measures, I argue that the process described by the MDE is primarily a measure of teacher effort, so I use the bonuses for evaluations as a measure of input based P4P.

One empirical challenge will be to disentangle the effect of the P4P bonuses from the effect of management practices. It may be that districts that put larger sums of money at stake for teacher or small team level goals, also implemented more robust JEPD programs. Likewise, it may be that districts that put larger sums of money at stake for evaluations, also implemented a more robust evaluation process. It seems plausible that districts that focus rewards on teacher or small team level goals would also focus on JEPD and those that focus rewards on evaluations would also focus on the Danielson framework, in which case it would be difficult to say whether an increase (or decrease) in student outcomes is due to the financial incentive or the management practice. Fortunately there are data on each Q-Comp component, so I am able to measure P4P apart from the supporting management components.

The data on Q-Comp come from two sources. First, when a district is approved, the state sends an official letter of acceptance and these letters are available on the MDE's website. The letters describe the district's Q-Comp plan in some detail. I coded these letters, paying particular attention to the dollars at stake for P4P bonuses. Variables indicate thousands of dollars at risk, i.e. $0.5 = \$500$ per teacher. Secondly, using a freedom of information act, I obtained data on the state's assessment of how well each district was doing with their management reforms. Specifically, the MDE conducted a formal review of each Q-Comp program in 2009. As part of the review, they scored each district on a rubric. Each section of the rubric focuses on one of Q-Comp's components and each section is comprised of sub-sections where a district can score "below proficient," "proficient" or "exemplary."[20] I construct a measure of the JEPD and evaluation components that measure

[20]To be exact, the rubric evaluates four of the five components. The alternative salary schedule component is not evaluated. It appears that many districts have struggled with this and that the MDE has

the percent proficient on the relevant rubric section. More detail on the rubric is provided in the appendix.

Summary statistics of the percent proficient on each component along with dollars at risk for the various P4P bonuses are provided in Table 2.1 and histograms show the distribution of each. Additionally, Table 2.2 shows the correlation between the various contract components. I also include summary statistics and correlations for two interaction terms that are used in the models described in the following section. There is a strong negative correlation between teacher or small team level P4P, $TeacherP4P\$$, and evaluation based P4P, $EvaluationP4P\$$. Districts appear to have chosen either to emphasize output based P4P or input based P4P. It is also noteworthy that there are strong positive correlations between $TeacherP4P\$$ and its interaction with $JEPDScore$ and $EvaluationP4P\$$ and its interaction with $EvalScore$. This proves to be one problem for estimation of complementaries of P4P and supporting management reforms and is discussed at more length in the results section.

In sum, I am able to characterize each Q-Comp district's program in two important ways: (1) Dollars at stake for teacher or small team-level goals, school or district-level goals, and formal evaluations; (2) percent proficient on the MDE rubric for JEPD and the teacher evaluation process. JEPD clearly supports teacher or small team level P4P and the evaluation process clearly supports evaluation based P4P. There is no district management reform that supports school level P4P.

Additionally, I have data on when districts first applied to be part of Q-Comp, when they were approved for the program and, if applicable, when they withdrew from the program. Only a few districts withdrew from the program and the data are coded accordingly. These data are combined with a panel of student achievement and demographic data.

The achievement data from the MDE are the Minnesota Comprehensive Assessment tests (MCA). Additionally, I use student achievement data from the Northwest Evaluation Association's Measures of Academic Progress (NWEA). Both the MCA and the NWEA have reading and math tests. The MCA is state-mandated. Prior to 2005 it was administered to all students in Minnesota public schools in grades 3, 5, and 7. Starting in 2005, it as administered to all grades 3-8. The NWEA is an additional test that some districts

deemphasized this component.

Table 2.1: Descriptive Statistics of Q-Comp Contract Elements

Variable	Mean	Std. Dev.	Min.	Max.	N
Teacher P4P$	0.84	0.694	0	2.2	67
School P4P$	0.244	0.205	0	0.8	67
Evaluation P4P$	1.101	0.700	0	2.5	67
JEPD Score	0.822	0.175	0	1	67
Eval Score	0.727	0.182	0.125	1	67
(Teacher P4P$)*(JEPD Score)	0.661	0.557	0	2.2	67
(Evaluation P4P$)*(Eval Score)	0.809	0.566	0	1.667	67

P4P$ variables are measured in $1,000. JEPD and Eval score are measured as percent proficient on the state's rubric. Each variable is weighted by the number of MCA-tested post-adoption student-years.

choose to purchase and use for diagnostic and other purposes. The sample of districts is thus smaller for the NWEA, however, having a second achievement measure is desirable because it provides insight as to whether P4P and associated reforms are inducing teaching to a specific test or more meaningful changes in instruction (Neal, 2011). For both tests, scores are standardized to mean 0 standard deviation 1 to facilitate interpretation and pooling across grades.

Demographic data for students who took each test are summarized in Table 2.3.

The resulting panel data provides a unique opportunity to investigate the effect of both input and output based P4P and complementary management reforms in education. I am aware of no other instance where there is such rich variation in P4P design coupled with data on management practices. The empirical results from this investigation have the potential to inform educational policy as well as provide insights for optimal contracting more generally.

2.5 Empirical Methodology

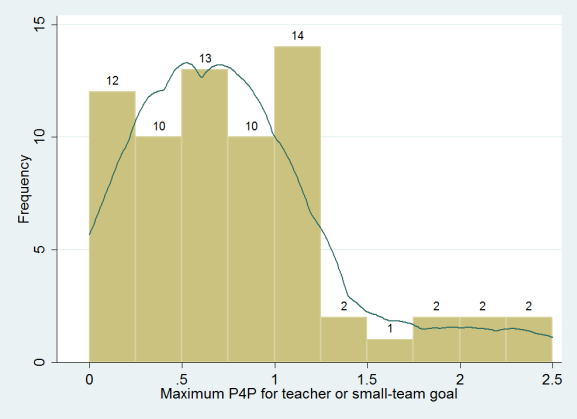
Building on Sojourner, West and Mykerezi (2011), I investigate how districts' student achievement changes as their Q-Comp participation changes and test whether districts that pair either input or output based P4P bonuses with complementary management

Table 2.2: Correlations Between Q-Comp Contract Elements

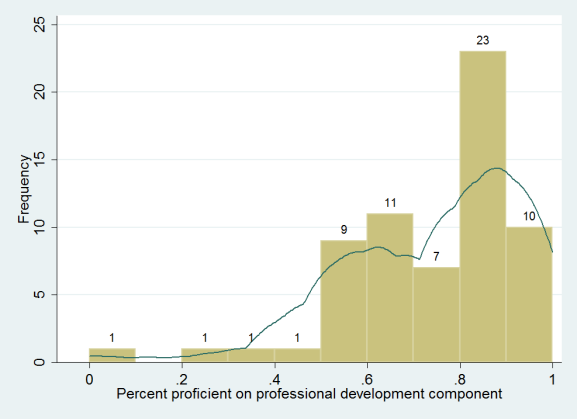
	Teacher$	School$	Eval$	JEPDScore	EvalScore	Teacher$*JEPD	Eval$*Eval
Teacher P4P$	1.0000						
School P4P$	0.0982	1.0000					
Evaluation P4P$	-0.7542	-0.1019	1.0000				
JEPD Score	-0.2474	-0.0629	0.2712	1.0000			
Eval Score	-0.1841	-0.0073	0.0715	0.4293	1.0000		
(Teacher$)*(JEPD Score)	0.9655	0.1212	-0.7001	-0.0341	-0.1126	1.0000	
(Eval$)*(Eval Score)	-0.7332	-0.0868	0.9142	0.4045	0.4229	-0.6707	1.0000

Figure 2.3: Distribution of P4P$ and Rubric Scores

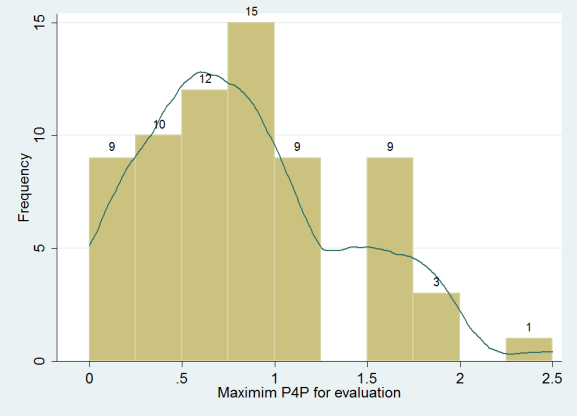
Teacher P4P



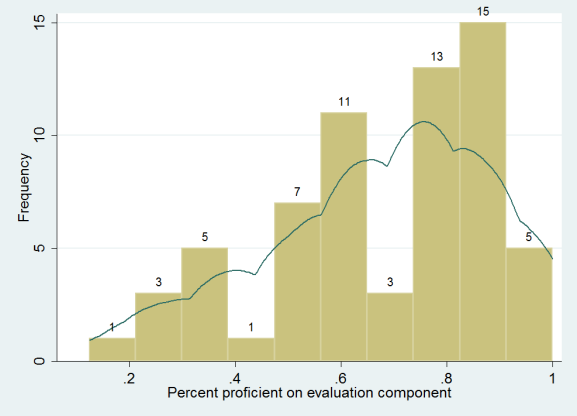
Job-Embedded Professional Development Score



Evaluation P4P



Evaluation Score



School P4P

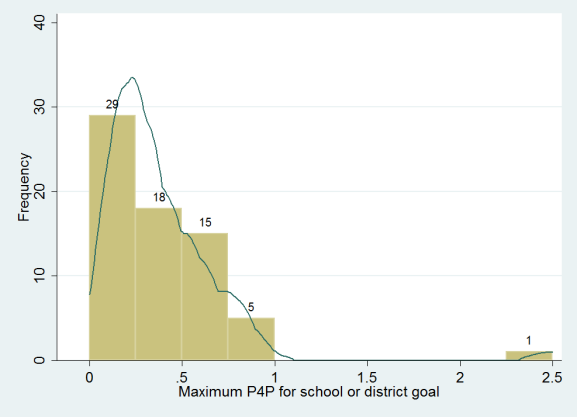


Table 2.3: Descriptive Statistics of Student Demographics for Q-Comp and Non Q-Comp Schools

Variable	Mean	Std. Dev.	Min.	Max.
MCA Sample	(2,052,337 student-years)			
Share male	0.513	0.06	0	1
Share free lunch	0.242	0.188	0	1
Share special ed.ucation	0.136	0.07	0	1
Share African-American	0.082	0.126	0	1
Share Hispanic	0.059	0.084	0	1
Share Asian-American	0.057	0.084	0	1
Share Native American	0.021	0.072	0	1
Enrollment (school-grade-year)	170.7	139.0	1	826
NWEA sample	(651,891 student-years)			
Share male	0.512	0.052	0	1
Share free lunch	0.197	0.119	0	1
Share special education	0.133	0.052	0	1
Share African-American	0.037	0.051	0	1
Share Hispanic	0.049	0.062	0	1
Share Asian-American	0.035	0.041	0	1
Share Native American	0.023	0.079	0	1
Enrollment (school-grade-year)	193.2	152	1	826

practices raise student achievement more than schools that do not. The outcome, y_{igdt}, is achievement for student i on the MCA or NWEA for each grade $g = 3...8$ in district d in year t. Scores are standardized within grade-year to mean zero and standard deviation one. I report results for math and reading separately using the following specification:

$$
\begin{aligned}
y_{igdt} \quad = \quad & \beta_1 QComp_{dgt} && (2.3)\\
& + \beta_2 TeacherP4P\$_{dgt} + \beta_3 SchoolP4P\$_{dgt} + \beta_4 EvaluationP4P\$_{dgt} && (2.4)\\
& + \beta_5 JEPDScore_{dgt} + \beta_6 EvalScore_{dgt} && (2.5)\\
& + \beta_7 (TeacherP4P\$_{dgt} * JEPDScore_{dgt}) && (2.6)\\
& + \beta_8 (EvalP4P\$_{dgt} * EvalScore_{dgt}) && (2.7)\\
& + \varphi 1(Dropped_{dgt}) + \alpha w_{sgt} + \gamma_i + \delta_t + \epsilon_{it} && (2.8)
\end{aligned}
$$

The first independent variable is an indicator for Q-Comp participation. Next, I include three variables that describe Q-Comp districts' P4P according to the maximum bonus available for teacher or small-team level goals, $TeacherP4P\$$, school or district level goals, $SchoolP4P\$$ and formal evaluations, $EvaluationP4P\$$. I describe Q-Comp districts' management components using the percent proficient on the state's rubric for the job-embedded professional development component, $JEPDScore$, and the evaluation component, $EvalScore$. I test the complementarity of these management practices with $TeacherP4P\$$ and $EvaluationP4P\$$ respectively.[21] There is no corresponding management reform that supports $SchoolP4P\$$.[22]

[21] Until now, I have largely focused on the importance of management practices that support delegated decision making, in the case of Q-Comp this is JEPD. The theoretical model also draws attention to the importance of observable teacher effort. Including information in the empirical model about bonuses for evaluations provides the opportunity to evaluate the impact of rewarding observable effort and/or implementing management policies aimed at measuring observable effort. That is, I can ask whether input based contracts work in education (where inputs are defined by evaluations rather than the traditional practice of rewarding degrees and years of experience) and whether management reforms intended to support the measurement of inputs complement this type of compensation.

[22] The interaction terms included in the model are the ones that are most theoretically relevant. For instance, it is not expected that $EvaluationP4P\$$ will interact in a meaningful way with $JEPDScore$. Table 7.2 in the Appendix empirically tests the importance of all possible interaction terms. F-tests of the joint significance of *all* the possible interaction terms reject the null hypothesis that they are statistically insignificant for all four measures of student achievement. F-tests of the joint significance of the interaction terms *excluded* from the main specification show that, for three of the four measures of student achievement,

Each of these variables is indexed by district-grade-year, however, in most cases Q-Comp plans are the same across all schools in participating district-years. The few exceptions were coded appropriately.

When controls are included for P4P and management reforms, β_1 measures the effect of implementing Q-Comp without any P4P or management reforms. More precisely, it measures the impact of Q-Comp with \$0 in bonuses and 0% proficient on the state's rubric for the JEPD and evaluation. It is worth noting that this hypothetical contract is not observed in the data so the coefficient should be interpreted with care.

If $\beta_2 > 0$, $\beta_3 > 0$, and/or $\beta_4 > 0$ then teacher-level, school-level or evaluation based P4P lead to increased student achievement when implemented alone. If $\beta_5 > 0$ and/or $\beta_6 > 0$, job-embedded professional development and/or formal evaluations lead to increased student achievement when implemented alone.

The interaction terms are of particular interest. First, $TeacherP4P\$ * JEPDScore$ measures the joint impact of teacher or small-team level P4P and job embedded professional development. A positive coefficient indicates that the two are complements and are best implemented together. If the effect of rewarding teacher level goals is greater when the JEPD is stronger (or the effect of JEPD is greater when more dollars are at risk for teacher level goals) then β_7 will be positive. Second, $EvaluationP4P\$ * EvalScore$ measures the joint impact of rewarding evaluations and the strength of the district's management policies regarding evaluation procedures. A positive coefficient would indicate that the two are complements and best implemented together. If the effect of rewarding evaluations is greater when the management practices surrounding the evaluation are stronger (or the effect of the management practices is greater when more dollars are at risk for the result of the evaluation), then β_8 will be positive.

Since Q-Comp participation is not randomly assigned, there may be systematic differences between districts that influence both Q-Comp adoption and student achievement, which would bias the estimates. To guard against this, I include the school-grade-year student demographics (w_{sgt}) reported in Table 2.3 to control for time-varying observable differences in peer characteristics and student fixed effects (1_i) to control for time-invariant

they are statistically insignificant. For NWEA reading the F-test rejects the null hypothesis that they are jointly insignificant. This almost complete lack of empirical evidence for the inclusion of these interaction terms supports the theoretical reasons for excluding them from the main specification.

unobservable differences in student characteristics. The model is identified from within-district-cohort, across-time variation. Fixed effects for each year (1_t) are also included. These terms identify counter-factual year effects. This is a generalization of difference-in-difference analysis that relies on differences in the timing of adoption across districts to separate time effects from program effects.[23] I also include an indicator for districts that were in Q-Comp but have since dropped out, $1(Dropped_{dt})$.

Lastly, I test for the importance of pre-adoption trends that might bias the coefficients of interest using an indicator for academic years two or more years prior to adoption, $(Pre-adoption_{dt})$ as well as this indicator interacted with the relevant contract elements. These results are presented in Appendix Tables 7.3 and 7.4. In this case, the reference category is the single year immediately prior to adoption. A negative coefficient on the pre-adoption indicator would mean that Q-Comp districts were improving prior to adoption and a positive coefficient would indicate that Q-Comp districts were declining prior to adoption (Lovenheim, 2009a; Sojourner et al., 2012). To attribute gains in achievement to Q-Comp, the coefficient on pre-adoption needs to be zero. If it is non-zero, we worry that districts select into Q-Comp base on achievement trends.

All standard errors are corrected for heteroskedasticity and correlation within district.

2.6 Results

MCA Results. The results with district-grade-year scores on the MCA reading exams as the outcome are reported in Table 2.4. Observable peer demographics (w_{sgt}) and indicators for dropping Q-Comp and days elapsed between tests are excluded from the table but are of expected sign and are reported in Appendix Table 7.1.[24]

The first column reports results for a regression that excludes all descriptors of the

[23] The first difference is the within-student comparison across time periods. The second difference is between the first-differences for students in adopting districts and those in non-adopting districts across the same time period. A within-student change between any two points in time is evaluated against changes across those same two years among other students.

[24] Readers may be surprised that the coefficient on share free lunch is positive. Recall that student fixed effects are included in the model so the student demographic variables account for peer effects, holding constant the student's own demographics. Thus, a positive coefficient on share free lunch is likely indicative of the fact schools with more students who are eligible for free lunch receive additional resources through the federal Title I program.

contract other than the simple indicator for Q-Comp participation. In this case we see that participating in Q-Comp is associated with a 0.031 standard deviation increase in reading scores that is statistically significant at the 5% level. Appendix Table 7.3 shows that models that include an indicator for two or more years pre-adoption find that this is not statistically significant. This suggests that Q-Comp adopters were no different from non-adopters prior to adoption and adds credibility to interpreting the coefficient on post-adoption as causal.

The second column reports results for a regression that introduces controls for P4P across the three dimensions - teacher or small team level rewards, rewards based on evaluations and school or district-level rewards. Focusing on $TeacherP4P$, the estimate of 0.019 indicates that for each $1,000 attached to individual or small-team level goals, reading scores increase by 0.019 standard deviations, although forcing a linear functional form may be problematic so the reader is cautioned against extrapolating this to conclude that $2,000 would yield a 0.038 standard deviation improvement and so on.[25] The fact that the post-adoption indicator is now negative (or at least non-positive) means that implementing Q-Comp without any P4P monies is not effective. The coefficients on the three different types of P4P are all between 0.019 and 0.041 and none are statistically significant suggesting that there is no clear "winner." The different types of P4P appear equally effective (or ineffective).

The third column reports results for a regression that adds controls for the management practices. Interestingly, scoring higher on the job-embedded professional development component of the state's rubric is negatively correlated with reading scores. Although statistically insignificant, this could be interpreted as evidence that this type of professional development is ineffective.

The fourth column reports results for a regression that includes the interactions to test for complementarities between P4P and management practices. The interaction terms change the interpretation of the other coefficients. For example, the coefficient on $JEPDScore$ is now the effect of going from 0 percent proficient to 100 percent proficient on the rubric when *no* money is tied to teacher-level goals (i.e. $TeacherP4P\$ = 0$) and the coefficient

[25] Note that the support from the sample is only in the range of $0 - $2,500.

on $TeacherP4P\$$ is the effect of an additional \$1,000 when the district scores below proficient on every element of the state's rubric for JEPD, (i.e. $JEPDScore = 0$). In this specification, the coefficient on $TeacherP4P\$$ is very small (and negative). This suggests that providing rewards for teacher or small-team level goals without also implementing job-embedded professional development is not effective. Put another way, the positive impact of $TeacherP4P\$$ in columns (2) and (3) is driven by districts that also implemented job-embedded professional development.

The positive coefficient on the interaction of $TeacherP4P$ and $JEPDScore$ seems to indicate that the two reforms are complements. This is not statistically significant at conventional levels, but it suggests that the impact of rewarding teacher-level goals may be increasing in the strength of the complementary professional development. This provides at least some support for the theory discussed in Section 2. I can not say, however, that districts that implemented teacher or small team level P4P and JEPD always improved reading scores. That fact that the coefficients in column 4 on $TeacherP4P\$$, $JEPDScore$ and $TeacherP4P\$xJEPDScore$ sum to -0.031 means that a district which put \$1,000 at risk for teacher or small team level goals and scored 100% proficient on the state's rubric for JEPD did not see an improvement in reading scores. If we assume a linear form for $TeacherP4P\$$ a district that put \$3,000 at risk for teacher or small team level goals would yield a positive result. This is a strong assumption, however, and a contract with \$3,000 for $TeacherP4P\$$ is not observed in the data. Further, An F-test of the joint significance of $TeacherP4P\$$, $JEPDScore$ and the interaction term fails to reject the null hypothesis that there is no impact of these variables on MCA Reading Achievement so, on the whole, the data proves inconclusive regarding support for the theory discussed in Section 2.

Surprisingly, when the interactions are included, coefficients on $EvaluationP4P\$$ and supporting management reforms are both large, positive and statistically significant but the interaction term is large, *negative* and statistically significant. This seems to indicate that monetary rewards for evaluation and evaluation procedures are substitutes rather than complements. This is surprising since presumably the two components of Q-Comp were intended to support each other. An F-test of the joint significance of $EvaluationP4P\$$, $EvalScore$ and the interaction term rejects the null hypothesis that there is no impact of these variables on MCA Reading achievement ($Prob > F = 0.0785$). The final column

includes only the interaction term and it is no longer negative (or large or statistically significant) casting doubt on result in column (4). Rewarding evaluations and/or providing management support for evaluations was not the focus of the theoretical model presented in Section 2 so I leave more extensive investigation of this issue for future research.

Results for MCA math achievement are in Table 2.5. Here there is no evidence that Q-Comp is an effective reform in the aggregate.[26] There is some evidence that evaluation based P4P is effective at raising math achievement. There is no evidence of complementarity between output based pay and support for decentralized decision making. The surprising result in column (4) indicating that evaluation based P4P and management support for evaluations are potentially substitutes reappears, however, when only the interaction term is included, column (5), this result again fades.

NWEA Results. Results using the NWEA test are presented in Tables 2.6 and 2.7. On the whole, Q-Comp appears effective at raising reading achievement on the order of 0.03 standard deviations and for the NWEA, there is evidence of a similar impact on math achievement. This result is strikingly similar to the result using the MCA test. This indicates that the gains from Q-Comp generalize to two different assessments and thus are less likely to be the result of "teaching to the test."[27]

In each table, the second column reports results for regressions that introduce controls for P4P across the three dimensions - teacher or small team level rewards, rewards based on evaluations and school or district-level rewards. Similar to the MCA, I find that rewarding evaluations appears most effective at raising NWEA math achievement. The coefficient of 0.074, from Table 2.7 column (2), indicates that a $1,000 bonus attached to subjective evaluations raises math achievement by 0.074 standard deviations. This results is robust to the inclusion of controls for management reforms in column (3). Table 2.7, column (3) indicates that job embedded professional development, as measured by $JEPDScore$, is not effective at raising math achievement, as measured by the NWEA tests.

In each table, column (4) reports results for regressions that tests the main hypothesis

[26] For math, as in reading, the indicators for two or more years prior to adoption and dropping Q-Comp are not significant, see appendix Table 7.3.

[27] See Sojourner et al. (2012) for more discussion of this.

Table 2.4: The Effect of Q-Comp on MCA Reading Achievement

	(1)	(2)	(3)	(4)	(5)
1(post-adoption)	0.031**	-.034	-.019	-.123	0.024
	(0.015)	(0.044)	(0.067)	(0.113)	(0.029)
Teacher P4P$		0.019	0.02	-.004	
		(0.022)	(0.021)	(0.102)	
School P4P$		0.041	0.041	0.037	
		(0.043)	(0.043)	(0.04)	
Evaluation P4P$		0.036	0.039*	0.131**	
		(0.023)	(0.023)	(0.055)	
JEPD Score			-.042	-.054	
			(0.07)	(0.106)	
Eval Score			0.019	0.193**	
			(0.048)	(0.088)	
(Teacher P4P$)*(JEPD Score)				0.027	-.002
				(0.136)	(0.023)
(Evaluation P4P$)*(Eval Score)				-.136*	0.011
				(0.072)	(0.026)
N Student-years	2,052,337	2,052,337	2,052,337	2,052,337	2,052,337
N Students	696,970	696,970	696,970	696,970	696,970
N Districts	369	369	369	369	369
Adjusted R^2	0.774	0.774	0.774	0.774	0.774

Dependent variable is student levels score on the MCA Reading Assessment normalized to mean 0 and standard deviation 1. Standard errors are corrected for heteroskedasticity and correlation within district. Significance *** 1%, ** 5%, * 10%. All specifications include year indicators, student fixed effects, and indicator for having dropped Q-Comp and covariates that measure school-grade-year student demographics.

Table 2.5: The Effect of Q-Comp on MCA Math Achievement

	(1)	(2)	(3)	(4)	(5)
1(post-adoption)	0.004	-.081	0.027	-.246*	0.041
	(0.021)	(0.055)	(0.104)	(0.129)	(0.056)
Teacher P4P$		0.031	0.024	0.043	
		(0.03)	(0.028)	(0.121)	
School P4P$		0.024	0.017	0.014	
		(0.065)	(0.06)	(0.056)	
Evaluation P4P$		0.048*	0.043	0.245***	
		(0.027)	(0.03)	(0.064)	
JEPD Score			-.103	-.075	
			(0.092)	(0.121)	
Eval Score			-.022	0.361***	
			(0.086)	(0.126)	
(Teacher P4P$)*(JEPD Score)				-.036	-.028
				(0.164)	(0.038)
(Evaluation P4P$)*(Eval Score)				-.303***	-.025
				(0.09)	(0.052)
N Student-years	2,007,029	2,007,029	2,007,029	2,007,029	2,007,029
N Students	686,484	686,484	686,484	686,484	686,484
N Districts	369	369	369	369	369
Adjusted R^2	0.792	0.792	0.793	0.793	0.792

Dependent variable is student levels score on the MCA Math Assessment normalized to mean 0 and standard deviation 1. Standard errors are corrected for heteroskedasticity and correlation within district. Significance *** 1%, ** 5%, * 10%. All specifications include year indicators, student fixed effects, and indicator for having dropped Q-Comp and covariates that measure school-grade-year student demographics.

of this paper, that P4P and management reforms are complements. Unfortunately, as with the MCA, the results are not clear. In the case of NWEA reading tests, it is again the interaction of $EvaluationP4P\$$ and $EvalScore$ that proves problematic. In the case of NWEA math tests, when the interaction of $TeacherP4P\$$ and $JEPDScore$ is included, the coefficients on each individually turn large and negative and the interaction term is large, positive and statistically significant and an F-test of the joint significance rejects the null hypothesis ($Prob > F = 0.0349$) of no effect. Although the theory in section 2 predicts a positive coefficient on the interaction term, the fact that both $TeacherP4P\$$ and $JEPDScore$ are large and negative means that, as with MCA reading, this is at best weak support for the theory since it is not the case that implementing both reforms together will necessarily lead to gains in student achievement. Indeed, when the interaction term is included without the individual terms, column (5), it is no longer significant.

Table 2.6: The Effect of Q-Comp on NWEA Reading Achievement

	(1)	(2)	(3)	(4)	(5)
1(post-adoption)	0.032**	-.004	-.016	-.036	0.015
	(0.016)	(0.044)	(0.067)	(0.082)	(0.035)
Teacher P4P$		0.024	0.026	-.076	
		(0.024)	(0.026)	(0.073)	
School P4P$		-.034	-.039	-.068	
		(0.047)	(0.044)	(0.043)	
Evaluation P4P$		0.021	0.018	0.087**	
		(0.024)	(0.026)	(0.044)	
JEPD Score			0.026	-.054	
			(0.063)	(0.082)	
Eval Score			-.007	0.146	
			(0.055)	(0.092)	
(Teacher P4P$)*(JEPD Score)				0.13	0.02
				(0.09)	(0.033)
(Evaluation P4P$)*(Eval Score)				-.110*	0.006
				(0.056)	(0.023)
N Student-years	651,891	651,891	651,891	651,891	651,891
N Students	247,026	247,026	247,026	247,026	247,026
N Districts	273	273	273	273	273
Adjusted R^2	0.793	0.793	0.793	0.793	0.793

Dependent variable is student levels score on the NWEA Reading Assessment normalized to mean 0 and standard deviation 1. Standard errors are corrected for heteroskedasticity and correlation within district. Significance *** 1%, ** 5%, * 10%. All specifications include year indicators, student fixed effects, and indicator for having dropped Q-Comp and covariates that measure school-grade-year student demographics and days elapsed between tests.

Table 2.7: The Effect of Q-Comp on NWEA Math Achievement

	(1)	(2)	(3)	(4)	(5)
1(post-adoption)	0.038	-.072	0.144	0.273	0.103
	(0.026)	(0.054)	(0.142)	(0.188)	(0.112)
Teacher P4P$		0.015	-.016	-.283	
		(0.043)	(0.044)	(0.184)	
School P4P$		0.026	0.057	-.006	
		(0.059)	(0.057)	(0.065)	
Evaluation P4P$		0.074***	0.079**	0.118	
		(0.025)	(0.031)	(0.093)	
JEPD Score			-.221**	-.462**	
			(0.11)	(0.185)	
Eval Score			-.057	0.052	
			(0.091)	(0.186)	
(Teacher P4P$)*(JEPD Score)				0.364*	-.061
				(0.213)	(0.088)
(Evaluation P4P$)*(Eval Score)				-.056	-.039
				(0.131)	(0.072)
N Student-years	655,341	655,341	655,341	655,341	655,341
N Students	247,768	247,768	247,768	247,768	247,768
N Districts	273	273	273	273	273
Adjusted R^2	0.838	0.838	0.838	0.838	0.838

Dependent variable is student levels score on the NWEA Math Assessment normalized to mean 0 and standard deviation 1. Standard errors are corrected for heteroskedasticity and correlation within district. Significance *** 1%, ** 5%, * 10%. All specifications include year indicators, student fixed effects, and indicator for having dropped Q-Comp and covariates that measure school-grade-year student demographics and days elapsed between tests.

Measurement Error and Estimation Challenges. In sum, strong empirical support for the complementarity of output based P4P and decentralized decision making is not evidenced using data from Q-Comp. This should not been interpreted as a repudiation of the theory, however, since measurement and estimation issues remain.

First, $JEPDScore$ is, at best, a noisy measure of support for decentralized decision making. The measure is based on the MDE's assessment of the "fidelity of implementation" of job-embedded professional development from a one time review of each district's Q-Comp reforms. The MDE conducted site visits and scored each district on a rubric that, while intended to be objective, likely depended on a fair amount of subjectivity. Further, the rubric was not designed for the purpose of measuring support for decentralized decision making, rather, I have appropriated a part of the rubric for this purpose. Additionally, the fact that I only have the $JEPDScore$ for one year, 2009, requires that I assume that districts were static in their adoption of this management practice. It is more likely that districts tinkered with their Q-Comp reforms over time as they engaged in and learned from an ongoing process of trial and error.

Second, the fact that including the interaction terms causes the coefficients on the individual components to change so dramatically suggests that there are underlying estimation issues. One possibility is that collinearity is plaguing the estimation. Both $TeacherP4P\$$ and $EvaluationP4P\$$ are highly correlated with their respective interactions with management reforms. To further investigate this issue, I transform the rubric scores for job embedded professional development and evaluation into binary rather than continuous variables. For these binary variables, a district that achieves at least 75% proficient is coded as a one, all other districts are coded as zero. I then interact the binary variables with the respective P4P$ components and use these interactions instead of the ones previously described. These results are presented in appendix Table 7.5. If it really were the case that the P4P$ and the reforms were complements or substitutes, we would expect to see this relationship hold despite this simple transformation. It does not. Instead, the interaction terms are no longer significant for all but one of the models this, along with the sum of the evidence discussed above, leads me to conclude that the large and statistically significant coefficients that sometimes appear on the interaction terms should be viewed with scepticism and not as support nor refutation of the theory.

2.7 Discussion

In this paper I use a principal-agent model adapted from Prendergast (2002). I offer evidence that education fits the assumptions of the model and an empirical test using data from Minnesota's Q-Comp program. I provide theoretical support for the complementarity of output based P4P and decentralized decision making but am unable to find empirical evidence due to measurement and estimation challenges.

I present a model, grounded in labor and personnel economics, that unifies the arguments for and against output based P4P for teachers. This model describes both the potential benefits and pitfalls of output based P4P and it incorporates a reality of educational production, namely the fact that teachers have important local knowledge, that is often ignored or only casually dealt with. I show that both the costs and benefits of output based P4P stem from the fact that teaching is a multidimensional work environment where there is considerable uncertainty. The main cost is lost productivity due to teaching to the test and the main benefit is ensuring that the lesson chosen is the one that best fits a particular set of students and a particular teacher.

Importantly, I identify the assumptions needed for output based P4P to work in education and suggest complementary management reforms. In order to realize the benefits of output based P4P, districts must delegate decision making responsibilities about curriculum and pedagogy to teachers. Additionally, districts may need to provide time and support for collaboration, to help teachers make sense of their local knowledge.

Opponents of output based P4P who are concerned with dangerous hidden actions will see these concerns accounted for in the model. Opponents who argue that output based P4P is demeaning to teachers should note the fact that the benefits of output based P4P in this model do not stem from the fact that teacher's are lazy or unmotivated. Teachers work hard in the absence of P4P, but they may focus their efforts on the wrong task. Output based P4P will work best when the district and individual teachers (or small teams of teachers) set a specific goal and the district provides an environment that fosters collaboration enabling teachers to make use of their local knowledge about the best way to achieve the goal in their particular classrooms. Proponents of output based P4P should advocate for complementary management reforms that give teachers more control over curriculum and pedagogy. Where teachers already have a good deal of control

over curriculum and pedagogy, reforms should focus on providing resources that support teachers in making good use of their local knowledge, i.e. time and opportunities for collaboration.

Too many policy debates lack any mention of collaboration. When they do, it is often opponents who argue that output based P4P will erode incentives to work with colleagues. I suggest that output based P4P may actually increase collaboration. If complementary management reforms are implemented alongside output based P4P, teachers will seek out more, not less, collaboration because collaboration is the mechanism via which teachers learn about the quality of the match between their specific classroom and the various lesson options. If teachers have preferences that are perfectly correlated with the district, it may be be enough to offer more opportunities for collaboration. I offer empirical evidence to the contrary. I find that time for collaboration alone does not increase student achievement. Nor does P4P alone. I find some evidence that the two are complements and are best implemented together but more work is needed. If further research confirms that the two are complements, an optimal contract will pair compensation reform and management practices that work in tandem. Policy makers should offer teachers more freedom to use their professional opinion about what is best, but with this must come some accountability for outcomes in the form of output based P4P.

Ideally future research could take the form of a randomized trial with three treatment groups. One group of districts/schools would adopt output based P4P without management reforms intended to support collaboration and another group would adopt the management reforms without the P4P. The third group would adopt both output based P4P and management reforms. Such a randomized trial is unlikely but existing policy variation may provide opportunities for natural experiments.

Although the data currently available from Q-Comp did not yield convincing empirical results, it may be possible to gather better data from Minnesota. First, the MDE is currently conducting a second program review and this will give a second set of rubric data. Since P4P and management reforms are likely evolving rather than static, rubric data from two points in time will provide a sense of how the reforms have changed over time and more accurately reflect what is happening in districts/schools each year. This is particularly important given that I rely on fixed effects models which exploit variation in

student achievement over time. Second, it is important to note that the MDE does not design its program review rubrics with this study in mind. It may be helpful to conduct an independent survey of participating districts to assess what exactly JEPD comprises in each district. It is very likely that JEPD looks different in each district and that some districts emphasize support for collaboration while others emphasize other things. It would also be helpful to survey non-participating districts to learn what sorts of support for collaboration exist in the absence of Q-Comp.

There are also opportunities to gather data from states other than Minnesota. For example, in Texas the District Awards for Teacher Excellence (DATE) program includes a wide variety of P4P plans with various supporting management changes.(Lincove, 2012) Researchers there indicate that there are potentially rich sources of data that could be exploited to test the theoretical model presented herein.

In sum, although the current study does not provide convincing empirical support for the theoretical model, there are existing and feasible extensions of the research that may provide a clearer picture. The research regarding P4P in education is notably lacking a discussion of the role that supporting management changes play and continuing in this vein of research will increase the chances that policy makers enact successful reforms. This research will also add to our knowledge about the complementarity of compensation and other aspects of job design more broadly and thus advance the fields of labor and personnel economics.

Chapter 3

Teachers' unions, compensation and tenure.

3.1 Introduction

In recent months governors and legislatures from states including Wisconsin, Ohio, New York and New Jersey have restricted collective bargaining rights for public employees, especially teachers. Increasingly, teachers' unions are blamed for inefficiencies and waste in the school system. Opponents argue that unions use their power to negotiate salaries that are not reflective of market forces and inflexible tenure policies that curtail districts' ability recruit, reward and retain the best teachers. This study investigates the validity of these claims by exploring the differences between teacher contracts in districts that are unionized and those that are not.

The standard contract for teachers is built around the single salary schedule, also know as "steps and lanes." Historically this contract is associated with unionization, however, it is also the dominant contract in non-unionized districts. With a steps and lanes contract, teachers earn additional compensation for experience, i.e. steps, and for college course work, i.e. lanes. All teachers with equal experience and education are paid the same regardless of other characteristics and teaching assignment. Unfortunately, research has shown that experience and education are poor proxies for teacher effectiveness.[1] Therefore

[1]See Hanushek (2003) for a review of this literature. Effectiveness improves in the first few years of

the single salary schedule restricts districts' ability to respond to labor market forces and does not allow districts to differentiate between effective and ineffective teachers.

Districts across the country are experimenting with contracts that depart from strict adherence to the single salary schedule. With the encouragement of national programs such as Race to the Top and the Teacher Incentive Fund, many states and districts have implemented pay for performance (P4P) and other reforms. In education, P4P can refer to policies that link teacher compensation to outputs, i.e. student test scores, or inputs, i.e. teacher actions. Although economists often assume that P4P refers only to output based rewards, the majority of policies in education are either entirely or at least partially input based. One major problem with previous research on the prevalence and determinants of P4P is that researchers have been unable to differentiate between different types of P4P. Input and output based P4P are very different and union support or opposition likely depends on exactly what measures of performance are rewarded.

Some districts, notably Washington D.C., have paired compensation reform with changes to the tenure system. Teachers traditionally earn tenure after three years of service to the district and once a teacher has tenure they can not be laid off or fired without cause. In D.C., teachers were offered the option to waive tenure protections in exchange for the opportunity to earn bonuses based on their measured effectiveness. The local teachers' union staunchly opposed these reforms and the superintendent who championed the reforms was forced out. The D.C. experience and other anecdotal evidence suggest that teachers' unions oppose compensation and tenure reforms but there is little formal research on the topic.

In this paper I systematically investigate the effect of teachers unions on compensation and tenure policies. I present new data from a survey of over 400 districts in which I poll human resource professionals about district practices. I pair the resulting data with data from the 2003-04 and 2007-08 waves of the Schools and Staffing Survey (SASS) and the Common Core of Data (CCD)to provide evidence on the prevalence and determinants of contract incentives in education. I pay particular attention to the differences between unionized and non-unionized districts including differentiating between unions that have legal collective bargaining rights and those that have only meet and confer privileges.[2]

teaching, but the marginal return to experience after three to five years is negligible. College course work in math and science appears to be beneficial, but the majority of teachers earn masters degrees in curriculum and pedagogy and these do not generally translate to improved student outcomes.

[2]The term "meet and confer" is borrowed from the SASS. Districts officials are asked: "Does the district

Using data from the SASS, I find that districts where contracts are negotiated via collective bargaining have higher starting salaries and larger step and lane increases. Salaries in districts where teachers are organized but have only meet and confer privileges rather than full collective bargaining rights are also higher than in non-union districts, although there is no evidence that meet and confer districts have larger step increases. Meet and confer agreements are not legally-binding. Therefore, I conclude that legal protections for collective bargaining have their largest impact on teacher contracts by negotiating returns to experience. I find that returns to experience are 26% greater in districts with collective bargaining than districts without collective bargaining.

Data from the SASS suggest that unionized districts are less likely to use P4P than their non-union counterparts; however, data from my survey present a more nuanced story. I find that districts where teachers are organized are no less likely than non-unionized districts to link compensation to student performance, but unionized districts are on the order of 3.5 to 8.5 percentage points less likely to base compensation on administrator review than their non-union counterparts. Lastly, using data from my survey, I find that unionized districts are 16.3 percentage points less likely to consider student performance and 7.1 percentage points less likely to consider peer reviews when granting tenure. These results are largely consistent with economic theory, empirical evidence from other sectors and previous work on teachers' unions.

3.2 Previous Literature

Classical economic theory treats unions as monopoly suppliers that use their market power to raise wages. Very simply, this leads to inefficiencies in production because resources are not allocated according to market signals. Freeman and Medoff (1984) term this the "monopoly face" of unions. They put forth an alternative theory which they refer to as the "collective voice face" of unions. In this case, they argue unions may actually *increase* efficiency by using their members' superior knowledge of the production function and desire for better working conditions to increase productivity.

---

have an agreement with a teachers' association or union for the purpose of meet-and-confer discussions or collective bargaining?" The question is followed with a clarification: "Meet-and-confer discussions are for the purpose of reaching non-legally-binding agreements. Collective bargaining agreements are legally binding."

In either case, unions are membership run organizations that seek to maximize the utility of their members. A median voter model suggests that unions protect the interests of the average member rather than the marginal worker. This leads to seniority provisions in union contracts. Therefore, we expect to find a union wage premium for senior teachers and more stringent tenure protections in unionized firms than in non-unionized firms. Previous empirical work has confirmed these theoretical predictions. Freeman and Medoff (1984) review the literature through the 1970s and find a union wage premium of approximately 15% in the private sector and 5% in the public sector. Blanchflower and Bryson (2007) find that since Freeman and Medoff's review, the union wage premium has declined in the private sector but risen substantially in the public sector both in absolute terms and relative to the private sector.

With regards to P4P, unions clearly support higher wages for there members. The median worker, however, is risk averse, so unions generally oppose compensation in the form of bonuses, profit-sharing or stock options (Kaufman, 2007; Budd, 2007). Since these types of contingent compensation are not frequently used in the public sector, empirical evidence on this point is drawn entirely from the private sector. Most relevant to the current study, Ittner and Larcker (2002) find that unions increase the use of worker-related measures of performance, i.e. input based P4P, relative to nonworker-related measures such as financial performance metrics, i.e. output based P4P.

It is worth noting that public sector unions differ from private sector unions in important ways (Freeman, 1986; Gunderson, 2007). Public sector unions can influence management practices through the political process. Freeman (1986) argues that public sector unions, unlike their private sector counterparts, will seek to increase employment by using their political power to raise the demand for public services. Also, by increasing demand and thus agency budgets, public sector unions may be able to avoid the difficult choice between layoffs and wage cuts. While private sector unions are associated with increased layoffs (since they prefer layoffs of the marginal member rather than wage cuts for the average member), public sector unions are associated with considerably lower layoff probabilities (Allen, 1988).

The rise in public sector unionism over the last few decades is arguably the most notable trend in the field of industrial relations and teachers unions have led the way. According

to the Bureau of Labor Statistics, education has the highest unionization rate amongst all occupations. Despite this, there is little recent research on the union effect on teacher contracts. Freeman (1986) summarizes the early literature on teachers' unions and reports a union wage premium of between 3% and 21%. These results are largely based on cross sectional studies. The most notable study since Freeman's review is Hoxby (1996). Hoxby compiles longitudinal data on teachers' unionization from the 1970s to the 1990s using the Census of Governments and finds that unions increase average teacher salaries by 5%. She concludes that this is likely the result of union rent-seeking rather than a sign that unions increase productivity. In contrast, Lovenheim (2009b) also uses longitudinal data from the 1970s to the 1990s but finds no impact of unions on average pay. He attributes the difference between his results and Hoxby's to measurement error in the Census of Governments.

I contribute to this literature in a number of ways. First, previous work largely focuses on average salaries. I am aware of few papers that consider the determinants of salaries in more detail and none that use recent data.[3] There is increased political momentum to limit collective bargaining rights and reform teacher compensation. I provide a detailed and updated understanding of the union effect on teachers' salaries, which will help inform this policy debate. Second, I investigate the union impact on P4P. With the exception of West and Mykerezi (2011), the few papers that attempt this use a very broad and problematic definition of P4P. Specifically, Ballou (2001); Belfield and Heywood (2008); Goldhaber et al. (2008) all discuss unions and P4P using a question in the SASS that asks if districts "reward excellence in teaching." West and Mykerezi (2011) have better data but are limited by a small sample. I collect and analyze detailed data to provide a more accurate picture of P4P in education. This adds important clarity to a contentious topic. Third, I consider important contract measures beyond compensation, notably tenure. Theory predicts that unions will oppose reforms that give managers more leeway in hiring and firing and it is often asserted that teachers' unions are a barrier to tenure reforms but I am aware of no empirical evidence that supports this point.

[3] Holmes (1979) provides early evidence that unions increase the return to both experience and education. West and Mykerezi (2011) find that this remains the case in 2003-04. Ballou and Podgursky (2002) do not distinguish between returns to experience and education but rather focus on the overall wage-tenure profile and conclude that wage growth over the first ten to 15 years is higher than can be justified by labor productivity gains.

The chapter proceeds as follows. In the next section I describe the data and empirical methodology. Section 3.4 presents results. In section 3.4.1 are results regarding unions and the single salary schedule. In section 3.4.2, I investigate unions and P4P. I use data from the SASS, including the "rewards for excellence" question, but argue that this is far too vague. I contribute more nuanced measures from my own survey. Section 3.4.3 does the same for tenure policies. Section 3.4.4 discusses robustness and alternative specifications. I conclude with a discussion of how the findings fit within the larger literature and policy debate.

3.3 Data and Methodology

The data for the study come from three sources: the Schools and Staffing Survey (SASS), the Common Core of Data (CCD), and a survey that I conducted. The SASS is a survey administered periodically by the U.S. Department of Education using a stratified probability sample design. The CCD is an annual census of all school districts maintained by the Department of Education.

I use the two most recent waves of the SASS. In 2003-04, 4,421 districts were included in the sample; in 2007-08, 4,618 districts were included in the sample. The SASS is not designed to be a longitudinal survey so the data is treated as a pooled cross-section rather than a panel. This yields 9,039 district-year observations. In my analysis, each district-year observation is weighted using the weights provided. This produces results that are representative of all public school districts in the country in 2003-04 and 2007-08.

The SASS questionnaire that is sent to district officials includes a variety of questions about the school district's demographics, policies and practices. Most relevant for the current study, both the 2003-04 and 2007-08 waves of the SASS ask respondents whether they use a salary schedule (96% respond that they do) and for a few points on the schedule. Specifically, the SASS has data on the starting salaries for teachers with a B.A. or M.A. as well as salaries for teachers with ten years of experience. I use these data to construct measures of the returns to experience and degree in each district. The SASS also has a question that ask respondents whether they reward teachers for national board certification, excellence in teaching, working in an undesirable location and/or working in a field that is experiencing shortages.

Researchers commonly use "rewards for excellence in teaching" to study P4P (Ballou, 2001; Belfield and Heywood, 2008; Goldhaber et al., 2008). It is easy to see, however, that this is an unsatisfactory measure of P4P for at least two reasons. (1) It is likely to be measured with a substantial amount of error; "excellence in teaching" is very vague, and the question leaves much to the discretion of the survey respondent. (2) It likely encompasses a wide range of P4P incentives including, but not limited to, rewards for student performance, peer evaluation, administrator evaluation or taking on leadership roles. It provides no distinction between output based P4P and input based P4P. West and Mykerezi (2011) find that unions are not monolithicaly opposed to P4P. More specifically, they find that unionized districts are less likely to reward teachers for student test scores, i.e. output based P4P, but more likely to reward teacher actions, i.e. input based P4P. These results, however, are based on a sample of only 100 districts.[4]

In order to better understand the prevalence and determinants of different types of P4P incentives and the union impact on the design of P4P, I conduct a more detailed survey. Between May 2009 and January 2011 research assistants under my supervision contacted human resource professionals via telephone and email to ask about district compensation and tenure policies. Initial contact was made using phone numbers from the CCD. If the person contacted preferred not to complete the survey by phone, a link was sent to an online survey instrument with the same questions. Districts were contacted in descending order based on total enrollment. Deviations were made to allow for time zone difference and other practicalities with one notable exception; districts in Minnesota were oversampled.[5] This was done to assist with a concurrent project that focuses on P4P in Minnesota.[6]

The sampling frame includes all districts with at least 7,000 students.[7] Not all districts responded. In some cases, the phone number in the CCD was no longer correct, so no contact was made. In other cases, the person contacted did not return calls or declined

[4]West and Mykerezi (2011) use data from the TR3 dataset compiled by the National Council on Teacher Quality. This dataset includes information on 100 of the largest districts in the nation. This data is reasonably detailed but not a representative sample. The largest district in each state is included and the remainder of the districts are the largest districts in the country. This method oversamples states such as Florida with large districts.

[5]Table 7.6 in the appendix shows the breakdown of respondents by state. A complete list of respondents is available upon request.

[6]Results excluding the districts in Minnesota that would not have been sampled based on total enrollment are available upon request and are nearly identical to those reported in this paper.

[7]There are some Minnesota districts with fewer than 7,000 students in the survey.

to participate. A few districts started the online survey but never completed the form. In the end, 1,228 districts were contacted and 410 completed the survey – a 33% response rate. Table 7.7 in the appendix shows the results for a probit model where the dependent variable is equal to one for a completed survey. Explanatory variables include all those listed in Table 3.1 along with state fixed effects and an indicator for charter schools status. The results show that districts that completed the survey are not different from districts that did not complete the survey along these, observable, dimensions (of course, districts that complete the survey may still be different from districts that did not complete the survey along unobservable dimensions.) An F-test of the joint significance of all of the variables with the exception of state fixed effects fails to reject the null hypothesis that these observable district characteristics do not predict survey completion ($Prob > chi2 = 0.3829$). For the state fixed effects, only Minnesota has a statistically significant higher response rate than the other states ($Prof > |z| = 0.021$).[8]

To further address concerns that response may be non-random, I construct a simple weighting scheme. I predict a completed survey response as a function of various district characteristics and use the inverse of the predicted probability of response to weight each observation.[9] I use these weights in an attempt to replicate an analysis that would represent the impact of unions on the compensation and tenure policies in all districts with at least 7,000 students.

One advantage of focusing on big districts is that they employ more teachers. Therefore, the resulting data describe contracts that cover a larger fraction of the nation's teachers than a random sample of all districts would. The 410 districts that completed the survey comprise just over 2% of all public school districts in the country but they employ 13.5%

[8] An F-test of the joint significance of the state fixed effects rejects the null hypothesis that state does not predict survey completion ($Prob > chi2 = 0.0000$). This is driven by Minnesota and the fact that DC and Hawaii have only one district – both of which were contacted but did not complete the survey and Maine, Montana, New Hampshire and Rhode Island where none of the districts contacted completed the survey. In Maine only one district was contacted. In Montana and New Hampshire only two districts were contacted. In Rhode Island five districts were contacted.

[9] Specifically, I use a probit model where the dependent variable is equal to one for a completed survey. Explanatory variables include per capita income in the district, the share of the district's population with a college degree, share of white students, share of African American students, share of Hispanic students, share of Asian students, share of students eligible for free and reduced lunch and indicators for state. All these are taken from the CCD and were chosen from all the available district characteristics using stepwise regression techniques. Weights were trimmed at the 5th and 95th percentile.

of public school teachers. I survey only 6% of districts in the 2007-08 SASS but, since they are disproportionately large districts, they employ 18.5% of teachers that work in SASS districts.[10] Table 3.1 compares the demographic characteristics of the districts that completed the survey to the CCD, a census of all districts, as well as districts sampled in the 2007-08 SASS. As expected based on the methodology, the districts in the survey are much larger than the average district as measured by total students, total staff (full time equivalents) and the number of schools. Also, since large districts are not a random sample, it is not surprising that these districts differ from the average in other ways. Most notably, they are less white and more Hispanic with more English language learner (ELL) students than the average district.

My survey includes detailed questions about district compensation and tenure policies. Table 3.2 summarizes the responses regarding compensation. The first column shows the share of all districts surveyed that base teacher compensation in various elements. The second column narrows the sample to the 282 districts that are also in the 2007-08 SASS. My survey was not intended to overlap directly with the SASS. However, the 282 districts that are in both my survey and the 2007-08 SASS are the focus of my analysis because I have the most information about these districts. Most importantly, for districts also in the SASS, I know unionization status. The next two columns show the break down of responses by unionization status and the final column reports the difference in means between these two groups.

These summary statistics show that unionized and non-unionized districts have very similar compensation policies. They also show that, despite recent reform efforts, basing pay on peer or administrator review or on student performance is still very uncommon, fewer than ten percent of districts use these types of P4P. The use of input based P4P is more common. Over 50% of unionized and non-unionized districts reward teachers for professional development. A simple comparison of means shows that unionized districts are less likely to base compensation on class assignment, administrator review or student performance than are non-unionized districts. Of course summary means may confound

[10]Of the 410 districts that completed the survey, 282 were also sampled in the 2007-08 SASS. Figure 7.1 in the appendix compares enrollments in these districts to all districts in the 2007-08 SASS with at least 7,000 students. Figure 7.1 shows that my sample is smaller than, but consistent with, the SASS. This supports my assertion that my survey is a representative sample of districts with at least 7,000 students.

Table 3.1: Demographic Characteristics of Districts Surveyed Compared to a Census of All Districts (CCD) and a Nationally Representative Sample (SASS)

	Survey mean N=410	CCD mean N=16,036	Difference Survey-CCD	SASS mean N=4,618	Difference Survey-SASS
Total students	17,800	2,881	+14,919***	6,865	+10,935***
Full time equivalent staff	1,052	179	+873***	429	+623***
Number of schools	27	5	+22***	12	+15***
Pupil-teacher ratio	17	15	+2***	15	+2***
Share pop college educ	0.15	0.13	+0.02***	0.13	+0.02***
Per capita income	21,426	19,694	+1,732***	19,441	+1,985***
Share pop below poverty	0.10	0.11	-0.01	0.12	-0.02***
Share male students	0.50	0.50	0.00	0.51	-0.01***
Share black students	0.12	0.11	+0.01	0.11	+0.01
Share white students	0.58	0.68	-0.10***	0.68	-0.10***
Share hispanic students	0.20	0.13	+0.07***	0.12	+0.08***
Share FRL students	0.43	0.46	-0.03***	0.46	-0.03***
Share ELL students	0.10	0.04	+0.06***	0.05	+0.05***
Share special education	0.13	0.14	-0.01	0.14	-0.01***

$***$ indicates significance at the 1% level.

Table 3.2: Survey Questions Regarding Compensation Policy - summary statistics

	All districts N=410	In SASS N=282	Unionized N=181	Non-Unionzed N=101	Difference Union-NonU
Years of experience	0.98	0.98	0.98	0.98	0.00
Level of education	0.98	0.98	0.98	0.98	0.00
Additional duties	0.96	0.95	0.94	0.97	-0.03
Specific classes	0.20	0.21	0.18	0.26	-0.08***
Professional development	0.52	0.53	0.52	0.56	-0.04
Peer review/observation	0.08	0.07	0.06	0.08	-0.02
Administrator review/observ	0.09	0.08	0.07	0.11	-0.04**
Student performance	0.07	0.07	0.05	0.09	-0.04**

** indicates significance at the 5% level. *** indicates significance at the 1% level.

the impact of unionization with other district characteristics.

Table 3.3 summarizes the factors considered when granting tenure. Like Table 3.2, it presents summary means for all districts that completed the survey and compares union and non-union districts for the 282 districts that are also in the 2007-08 SASS. The vast majority of districts consider administrator review when granting tenure. Complementary measures such as peer review or a teacher prepared portfolio are less common. Fewer than one in four districts consider student performance when granting tenure. Comparing across unionized and non-unionized districts, it appears that unionized districts are less likely to consider peer review and student performance when granting tenure. Again, this statement is based only on summary means and results that control for covariates follow shortly.

To further the analysis, I estimate versions of the following equation:

$$Contract_d = \beta_0 + \beta_1 CB_d + \beta_2 MC_d + \beta' X_d + \epsilon_d \tag{3.1}$$

All results in the next section are obtained from district-level estimates of equation (3.1). The dependent variable, $Contract$, describes some element of the contract in district d. CB_d is an indicator of the existence of a union that bargains collectively over the

Table 3.3: Survey Questions Regarding Tenure - summary statistics

	All districts N=410	In SASS N=282	Unionized N=181	Non-Unionized N=101	Difference Union-NonU
Administrator review/observ	0.90	0.92	0.93	0.90	0.03
Peer review/observ	0.15	0.13	0.10	0.18	-0.08**
Teacher portfolio	0.18	0.20	0.21	0.19	0.03
Student performance	0.22	0.22	0.17	0.31	-0.14***

** indicates significance at the 5% level. *** indicates significance at the 1% level.

contract, MC_d is an indicator of the existence of a teachers' union that has only meet and confer privileges. These categories are mutually exclusive. The omitted category is districts with no union.

The main coefficients of interest are β_1 and β_2. The cross-sectional nature of the data limit the modeling options and my ability to interpret β_1 and β_2 as causal. I use demographic controls, X, to account for observable district characteristics that may be correlated with both compensation and unionization but the lack of panel data prevents more convincing identification strategies.[11] Demographic characteristics included in X are total enrollment, racial composition of the student body, the share of students that receive free and reduced lunch, the total number of teachers, racial composition of the teachers, the share of school aged children in the district, the share of college educated residents in the district and per capita income. Variables not measured in the SASS are taken from the CCD. There are 8,294 district-years with data on the full complement of variables included in X.

In some models I include indicators for census region or state. Results are presented with and without these because region and state are highly correlated with unionization and

[11] Hoxby (1996) argues that union status should be instrumented. She finds that non-instrumental results are lower bounds of the true union effect. West and Mykerezi (2011) also find that non-instrumental results should be interpreted as lower bounds. In general, I find that the arguments for the endogenteity of unions and salary schedule characteristics not as convincing for my more recent data as they are for studies covering earlier eras. Specifically, union status is remarkably stable in recent years. There are almost no districts that switch from non-union to union and vice versa which mitigates concerns that compensation and/or tenure policies are causing unions to form rather than the reverse. This, and the lack of a convincing instrument lead me to prefer the simple non-instrumental estimates for this study.

thus raise concerns about multicollinearity with the variables of interest. The states where collective bargaining is illegal are exclusively in the south (although collective bargaining is not illegal in every southern state). When indicators for region are included, regional differences are likely confounded with the impact of unionization. When indicators for state are included, specific state policies are held constant but states where collective bargaining is illegal or universal do not contribute to the β_1 coefficient. Only states with some variation in bargaining status contribute to the estimate of the union premium in these models.[12]

I start with three measures of *Contract* to describe the basics of a district's salary schedule: (1) starting salary, (2) returns to experience and (3) returns to degree. Starting salary is measured in log form so β_1 is interpreted as a percent increase associated with the existence of collective bargaining relative to no union. Returns to experience and degree are measured as percent increases for an additional year of service or a masters degree, respectively, i.e. (salary with experience(degree) - salary without experience(degree))÷(salary without experience(degree)). For these, β_1 is interpreted as percentage point increases associated with the existence of collective bargaining relative to no union.

It may be the case that unions trade step increases for lane increases or vice versa or accept lower step and lane increases in exchange for higher starting salaries. If, for example, some local unions prefer larger step increases and others larger lane increases, the estimates of the union premium on these individually may be lower bounds. To account for this, I add two additional measures of *Contract* – the (log) salary of a teacher with a bachelors degree and ten years of experience and the (log) salary of a teacher with a masters degree and ten years of experience. The salary at year ten for a teacher with a masters degree has the advantage of combining the union effect on step and lane increases. By combining them, I am able to see the net effect of collective bargaining on salaries for midcareer teachers – likely the median union voter.

Next, I let *Contract* be indicators for the existence of rewards for national board

[12]Collective bargaining is illegal in Georgia, North Carolina, South Carolina, Texas and Virginia (National Center for Teacher Quality). With the exception if Virginia, there are effectively no unions in these states with meet and confer status either. Additionally, fewer than five percent of districts in Alabama, Arizona, Arkansas, Mississippi, Missouri, West Virginia and Wyoming are covered by collective bargaining agreements. However, some of these states allow meet and confer status (National Center for Education Statistics). Over 95% of districts in Hawaii, Iowa, and Nevada are covered by collective bargaining. When meet and confer unions are included, Connecticut, Florida, Indiana, Maryland, Oregon, Vermont and Wisconsin have over 95% union coverage (National Center for Education Statistics).

certification, excellence in teaching and teaching in a field experiencing shortages. These are the questions in the SASS that best approximate P4P policies. In this case the dependent variables are binary so I use probits. The results I present are marginal effects calculated at the mean.

The SASS does not have any data on tenure policies. However, there is information on the number of teachers who are fired. In both 2003-04 and 2007-08 the survey asks how many teachers "were dismissed or did not have their contracts renewed as a result of poor performance." The question refers to the previous school year. In 2003-04 respondents were asked to separate their response by teachers with three or fewer years of experience and teachers with more than three years of experience. In 2007-08 respondents were asked to separate their response by tenured teachers and non-tenured teachers.[13] To my knowledge these data have not been exploited by researchers, possibly because there are at least three hurdles to using these data. First, to pool the data, I refer to teachers with three or fewer years of experience and non-tenured teachers as "junior teachers" and teachers with three or more years of experience and tenured teachers as "senior teachers" but acknowledge that this combines two different definitions across years. Second, I have concerns about measurement error. The numbers reported for some districts seem unreasonably large – a few report dismissing over 80% of their teaching staff for poor performance. It is possible that these schools were restructured with a new staff but it is also possible that the respondents misunderstood the question.[14] Third, over half of the district report dismissing no teachers for poor performance.

Finally, I use the measures from my survey listed in Tables 3.2 and 3.3 for *Contract* and test the impact of unionization on the details of P4P and tenure policies. The advantage of the variables from my survey is that they provide more granularity with regard to the specific measures used when awarding either pay increases or tenure. This allows me to better understand what type of P4P unions oppose (or support). The majority of these are binary variables and results presented are marginal effects calculated at the mean from probit models. The only non-binary variable is the number of years before a teacher is

[13] Additional instructions directed districts without a tenure system to separate their responses by teachers on year-to-year contracts and permanent employees.

[14] To illustrate the impact of these outliers, note that topcoding the data at the 95th percentile drastically reduces the mean share of teachers dismissed from 3% to 1.4%.

granted tenure. For this, I use ordinary least squares.

In all models, standard errors are corrected for heteroskedasticity.

3.4 Results

3.4.1 Unions and the Single Salary Schedule

Table 3.4 summarizes the effect of collective bargaining on salary schedules. Column (1a) shows a raw correlation between the various measures of the single salary schedule and an indicator for collective bargaining, i.e. β_1 from an OLS estimation of equation (3.1) with no controls for district demographics. Column (1b) controls for total enrollment, racial composition of the student body, the share of students that receive free and reduced lunch, the total number of teachers, racial composition of the teachers, indicators for urban and rural locations, the share of school aged children in the district, the share of college educated residents in the district and per capita income. Column (1c) includes all of the aforementioned and indicators for census region. Lastly, column (1d) includes results with state indicators in place of indicators for census region.

Salaries are higher in districts with collective bargaining than in non-union districts. Salaries start higher and stay higher. The raw correlation in column (1a) shows that unionized districts have salaries that are 8% higher for new teachers and 17% higher for a teacher with a masters and ten years of experience than their non-union counterparts. The effect is smaller when controls are included but remains statistically significant in all specifications. The lower bound from column (1d), which includes state indicators, shows that, in states where teachers in some districts bargain collectively and others do not, salaries are three to four percent higher in districts with collective bargaining than in districts without a union.

The returns to experience and a masters are higher in districts with collective bargaining than in non-union districts. That is, districts with collective bargaining have larger step and lane increases than their non-union counterparts. The mean annual step increase in all districts is 2.7%. Column (1a) shows that in districts with collective bargaining, step increases are 0.86 percentage points higher. When demographic controls are included this

Table 3.4: The Effect of Collective Bargaining on Salary Schedules

	2003-04 and 2007-08 SASS			
Control variables	(1a)	(1b)	(1c)	(1d)
Demographic characteristics	no	yes	yes	yes
Census region fixed effects	no	no	yes	no
State fixed effects	no	no	no	yes
	N=8,294	N=8,294	N=8,294	N=8,294
Dependent variable	Effect of collective bargaining			
Starting salary log	0.0826***	0.0546***	0.0626***	0.0334**
	(0.0069)	(0.0095)	(0.0131)	(0.0163)
Returns to experience annual salary increase/base salary	0.0086***	0.0070***	0.0050***	0.0013
	(0.0005)	(0.0005)	(0.0007)	(0.0012)
Returns to degree salary increase for MA/base salary	0.0355***	0.0396***	0.0187***	0.0053
	(0.0031)	(0.0035)	(0.0044)	(0.0080)
Salary for BA + 10 yrs log	0.144***	0.103***	0.092***	0.037**
	(0.006)	(0.008)	(0.011)	(0.017)
Salary for MA + 10 yrs log	0.174***	0.135***	0.105***	0.038**
	(0.006)	(0.008)	(0.010)	(0.015)

Table reports β_1 from OLS estimations of equation (3.1) with various dependent and control variables. Robust standard errors are in parentheses. Significance: $*$: 10% $**$: 5% $***$: 1%. Demographic characteristics are total enrollment, racial composition of the student body, the share of students that receive free and reduced lunch, the total number of teachers, racial composition of the teachers, the share of school aged children in the district, the share of college educated residents in the district and per capita income.

drops to 0.70 percentage points – this is a 26% union premium on the returns to experience. The result is smaller, 0.50 percentage points, but remains statistically significant when census region indicators are included. The mean lane increase in all districts is 11%. Column (1a) shows that teachers in districts with collective bargaining receive an additional 3.55 percentage point increase for a masters. When demographic controls are included, the result is largely unchanged. When region is account for, the result is smaller, 1.87 percentage points, but still statistically significant. The union premium on step and lane increases is not robust to the inclusion of state indicators, column (1d). This is not entirely surprising since in some states, step and lane increases are set centrally and thus do not vary across districts within states.[15]

Another reason that the inclusion of state fixed effects likely moderates all of the results is that teachers are licensed by states and often these licenses do not easily transfer across state lines. Thus, districts only have to compete for teachers within state. In state variation in compensation is likely less than between state variation in compensation because non-union districts in states that permit unionization have to compete for teachers with unionized districts.

Table 3.5 summarizes the effect of meet and confer status on salary schedules, i.e. β_2 from an OLS estimation of equation (3.1). By comparing the effect of collective bargaining to the effect of meet and confer I am able to narrow in on the impact that limiting collective bargaining rights might have. Presumably, in states that pass laws limiting collective bargaining, teachers will remain organized but the results of their negotiations with the district will no longer be legally binding. By looking at the experience of meet and confer districts, I estimate the impact of unions without legally binding negotiations.

I find that in districts where teachers are organized but only have meet and confer status, salaries are higher than in districts without any union. The magnitude of the effect on starting salaries is somewhat less than for districts with collective bargaining in column (1a), but once controls are introduced, the magnitudes in columns (1b), (1c) and (1d) are strikingly similar. Indeed, the results for meet and confer are in some cases stronger than for collective bargaining. This suggests that, even without formal collective bargaining

[15] According to data from the National Council on Teacher Quality, there are currently 16 states (and the District of Columbia) that have a single salary schedule for the state.

Table 3.5: The Effect of Meet and Confer on Salary Schedules

	2003-4 and 2007-08 SASS			
Control variables	(1a)	(1b)	(1c)	(1d)
Demographic characteristics	no	yes	yes	yes
Census region fixed effects	no	no	yes	no
State fixed effects	no	no	no	yes
	N=8,294	N=8,294	N=8,294	N=8,294
Dependent variable	Effect of collective bargaining			
Starting salary	0.0583***	0.0501***	0.0643***	0.0480***
log	(0.0099)	(0.0097)	(0.0110)	(0.0109)
Returns to experience	0.0006	-0.0001	-0.0016**	0.0005
annual salary increase/base salary	(0.0008)	(0.0007)	(0.0008)	(0.0009)
Returns to degree	0.0299***	0.0345***	0.0152**	0.0063
salary increase for MA/base salary	(0.0059)	(0.0062)	(0.0058)	(0.0061)
Salary for BA + 10 yrs	0.0589***	0.0438***	0.0436***	0.0493***
log	(0.0106)	(0.0098)	(0.0106)	(0.0114)
Salary for MA + 10 yrs	0.0838***	0.0720***	0.0546***	0.0525**
log	(0.0105)	(0.0096)	(0.0109)	(0.0111)

Table reports β_2 from OLS estimations of equation (3.1) with various dependent and control variables. Robust standard errors are in parentheses. Significance: $*$: 10% $**$: 5% $***$: 1%. Demographic characteristics are total enrollment, racial composition of the student body, the share of students that receive free and reduced lunch, the total number of teachers, racial composition of the teachers, the share of school aged children in the district, the share of college educated residents in the district and per capita income.

rights, if teachers are organized, they are able to negotiate for higher starting salaries. The impact of meet and confer on starting salaries ranges from 4.8 to 6.4% depending on which controls are included.

After ten years, teachers in meet and confer districts earn between 4.3 and 8.3% percent more than their counterparts in non-unionized districts. In the models without state indicators, the magnitude of the effect of meet and confer at year ten is roughly half that of collective bargaining. However, in the model with state indicators, meet and confer districts seem to fair slightly better than collective bargaining districts.

In stark contrast to collective bargaining, districts with only meet and confer privileges do not have higher returns to experience, i.e. step increases. They do, however, have very similar returns to a degree, i.e. lane increases. This leads me to conclude that collective bargaining has its largest impact through negotiating larger returns to experience. As I discuss further in the conclusion section, this is consistent with economic theory and other empirical findings.

3.4.2 Unions and Pay for Performance

Table 3.6 summarizes the effect of collective bargaining on pay for performance. I report the marginal effect calculated at the mean from probit estimates of β_1 from equation (3.1) for various dependent variables. I use data from the 2003-04 and 2007-08 SASS and my survey. The first two columns use all the observations in the SASS while the second two columns use only those districts that are in both my survey and the SASS. The columns labeled with (b) have controls for demographics. The columns labeled (c) have controls for demographics and census region and those labeled (d) have state fixed effects in place of census region (the lettering references the fact that these are analogous to columns (b), (c) and (d) in previous tables.) The results in columns (d) should be interpreted with care. The state fixed effects are particularly problematic for models using data from my survey because there were often only a few districts sampled in each state. In fact, some states had only one district that completed the survey. Notice that the sample size drops when state fixed effects are included. For this reason, I focus my discussion on columns (b) and (c).

Table 3.6: The Effect of Collective Bargaining on Pay for Performance

	SASS			Survey		
Control variables	(2b)	(2c)	(2d)	(3b)	(3c)	(3d)
Demographic characteristics	yes	yes	yes	yes	yes	yes
Census region fixed effects	no	yes	no	no	yes	no
State fixed effects	no	no	yes	no	no	yes
	N=8,294	N=8,294	N=8,294	N=282	N=282	$N \leq$282
Dependent variable	Effect of collective bargaining					
Reward shortage field	-0.053***	0.007	0.034*	-0.137	-0.102	-0.327
	(0.014)	(0.013)	(0.018)	(0.084)	(0.103)	(0.203)
Specific classes				-0.243***	-0.067	0.163
				(0.082)	(0.087)	(0.223)
Additional duties				-0.008	-0.014	-0.975***
				(0.015)	(0.015)	(0.0388)
Reward national board	-0.072***	0.076**	0.043	0.028	0.217*	0.622***
	(0.016)	(0.023)	(0.043)	(0.094)	(0.120)	(0.213)
Professional development				0.015	-0.007	-0.262
				(0.089)	(0.109)	(0.189)
Reward excellence	-0.031***	-0.005	-0.001	-0.092	-0.060	-0.307**
	(0.007)	(0.008)	(0.007)	(0.063)	(0.076)	(0.154)
Peer review				-0.027	-0.028	.
				(0.024)	(0.030)	.
Administrator review				-0.063*	-0.085*	-0.116
				(0.033)	(0.048)	(0.150)
Student performance				-0.030	-0.034	-0.201**
				(0.033)	(0.035)	(0.093)

Table reports marginal effects of β_1 from probit estimations of equation (3.1) with various dependent and control variables. Robust standard errors are in parentheses. Significance: $*$: 10% $**$: 5% $***$: 1%. Demographic characteristics are total enrollment, racial composition of the student body, the share of students that receive free and reduced lunch, the total number of teachers, racial composition of the teachers, the share of school aged children in the district, the share of college educated residents in the district and per capita income. Columns (2b), (2c) and (2d) use the weights provided with the SASS. Columns (3b), (3c) and (3d) use the survey weights described in the text.

Results from the SASS indicate that districts where teachers bargain collectively are less likely to have any type of P4P, be it rewards for shortage fields, national board certification or excellence in teaching than non-union districts. These results, however, are not robust to the inclusion of census region (nor state indicators, results not shown).[16] Columns (3b) and (3c) show that districts in my survey with collective bargaining are less likely to attach rewards to shortage fields and excellence but more likely to reward national board certification than the non-unionized districts in my survey. The results in columns (3b) and (3c) are not generally statistically significant but the sample size is dramatically smaller than in columns (2b) and (2c).

My survey is designed to probe P4P policies in more detail than the broadly worded questions in the SASS. Table 3.6 is laid out to group like questions. For instance, I expect that a district which reports that they reward teaching in shortage fields in the 2007-08 SASS, will also report that they link compensation to teaching specific classes in my survey.[17] Indeed, I find that districts with collective bargaining are less likely to reward teaching in a shortage field and less likely to link compensation to teaching specific classes. This is consistent with theory that predicts a union "solidarity wage" which does not distinguish between workers based on job assignment.

I find that districts with collective bargaining are just as likely to pay for professional development as are non-union districts. This suggests that unions do not oppose differentiating between teachers based on specific training. This is consistent with the finding that unions support rewards for teachers who earn a Masters. Unions appear to support P4P that is based on measures of teacher inputs, be they the traditional measures of degrees or alternative measures such as district specific professional development.

Most interesting is the detail on output based P4P that my survey provides. Union opposition to pay for excellence does not seem to be concentrated in opposition to paying teachers for student performance. I find stronger evidence for union opposition to linking compensation to administrator review. Unionized districts are 6.3 to 8.5% less likely to tie pay to administrator review than non-union districts, depending on the specification

[16] The result for rewards for national board certification remains significant but switches sign.

[17] I do not expect the results to be exactly comparable since the questions were worded differently, the respondents were likely a different individuals and there may have been as much as three years between the SASS and my survey during which policies may have changed.

choice.[18]

Table 3.7 shows the full results for all the marginal effects calculated at the mean from probit estimates of each control variable in equation (3.1) with administrator review as the dependent variable. Note that meet and confer districts are also less likely to base compensation on administrator review than are entirely non-union districts. This suggests that formal collective bargaining is not the only road block to this type of compensation reform. Anywhere teachers are organized, districts are less likely to link pay to administrator review regardless of the legal standing of the union. The magnitudes of the coefficient on the collective bargaining are larger than the coefficient on meet and confer in all models which indicates that, where teachers have full collective bargaining rights, they are more effective in keeping out this type of reform. However, the magnitudes are small in either case. Unionized districts of either type are only between 3.5 and 8.5% less likely to link compensation to administrator review than are non-unionized districts.[19]

3.4.3 Unions and Tenure Policy

Lastly, Table 3.8 summarizes the effect of collective bargaining on tenure and dismissal. The layout of the table is as before – the first three columns use all the districts in the SASS and the last three columns focus on districts in both the 2007-08 SASS and my survey. I find that districts with collective bargaining are 7.5 to 10% more likely to grant tenure depending on the specification, although this result is not statistically significant when region controls are excluded. This magnitude is smaller than I suspect many union opponents would expect. Tenure, like the single salary schedule, is associated with unions but is in fact dominant in all public schools. Perhaps more in keeping with expectations, districts with collective bargaining grant tenure protections 0.74 to 0.83 years earlier than their non-union counterparts depending on specification choice.[20] Looking at what factors

[18] The statistical significance for this result is stronger for the probit coefficients rather than the marginal effects.

[19] Table 3.7 also shows that a larger share of minority students decreases the likelihood that a district will link teacher pay to administrator review. Although a larger share of African American teachers increases the likelihood of such a policy. I offer no theory for why this might be but it may be of interest for future research.

[20] As before, I focus on columns (b) and (c) because the state fixed effects in column (d) are problematic for my small survey.

Table 3.7: Full Results for Administrator Review

Control variables	(3a)	(3b)	(3c)	(3d)
Collective bargaining	-0.0738	-0.0633*	-0.0856*	-0.116
	(0.0493)	(0.0333)	(0.0483)	(0.150)
Meet and confer	-0.0649**	-0.0356***	-0.0355***	-0.0192
	(0.0329)	(0.0138)	(0.0133)	(0.0277)
Share Hispanic students		-0.273***	-0.237***	-0.110
		(0.0886)	(0.0850)	(0.196)
Share African American students		-0.458***	-0.424***	-0.269
		(0.144)	(0.136)	(0.320)
Share free/reduced lunch		0.0993	0.0938	-0.474
		(0.0663)	(0.0607)	(0.0710)
Share Hispanic teachers		-0.0511	-0.0255	-0.0680
		(0.117)	(0.0883)	(0.205)
Share African American teachers		0.351**	0.341**	0.286
		(0.156)	(0.143)	(0.333)
Share new teachers		0.111	0.171	0.0336
		(0.191)	(0.165)	(0.126)
City		0.0452	0.0357	0.0239
		(0.0439)	(0.0362)	(0.0452)
Suburb		0.0745	0.0766	0.103
		(0.0637)	(0.0594)	(0.122)
Rural		0.00957	0.0144	-0.0089
		(0.0441)	(0.0422)	(0.0161)
Share school aged population		0.385	0.302	0.174
		(0.326)	(0.308)	(0.300)
Share college educated population		0.0658	0.125	-0.170
		(0.274)	(0.244)	(0.343)
Per capita income		2.37e-07	-7.14e-07	7.24e-07
		(2.88e-06)	(2.52e-06)	(3.09e-06)
Observations	282	282	282	164
Psuedo R^2	0.028	0.247	0.265	0.397

Table reports marginal effects of β_1 from probit estimations of equation (3.1) with a dependent that is equal to 1 if a district reports that they base pay at least in part on administrator review or observation. Robust standard errors are in parentheses. Significance: $*$: 10% $**$: 5% $***$: 1%.

are considered when granting tenure, there is some evidence that districts with collective bargaining are less likely to include peer review and/or student performance when awarding this level of job security. In sum, tenure protections do not differ dramatically between unionized and non-union districts with the exception of the fact that districts with collective bargaining grant tenure almost one year sooner than their non-union counterparts.

Finally, one counter-intuitive finding that raises questions for future research is that districts with collective bargaining in my survey dismiss slightly *more* teachers for poor performance than non-union districts. This is entirely opposite conventional wisdom which asserts unions stand in the way of district's ability to get rid of ineffective teachers. While I do not have a hypothesis for this yet and the finding requires further scrutiny, I can say that at first blush, I do not find support for this criticism of unions.[21]

3.4.4 Robustness and Alternative Specifications

While thus far I have based my discussion on simple OLS and probit estimates, a few specification issues merit further attention. First, it is possible that unionization may be endogenous either due to reverse causality or omitted variables. For instance, teachers may choose to unionize based on low pay or, as Hoxby (1996) argues, there may be omitted variables such as poor management that lead to both low pay and unionization. In either case, simple OLS and probit estimates may be biased downward. Indeed, Hoxby's instrumental variable (IV) estimates, which control for endogeneity, produce somewhat higher estimates of the union effect on wages than her OLS estimates.[22]

To test whether my findings are biased I estimate several IV models using instruments that measure state laws granting teachers unionization rights, the percent of workers in all professions in the state who are members of unions in 1964 and the percent of immigrants in the state at the turn of the 20th century. I briefly describe the construction and rationale for each below and I report IV results in Appendix Table 7.8. In brief, like Hoxby (1996), I find that IV estimates indicate a larger impact of unions on compensation than OLS estimates.

[21] One possible explanation could be that unions do not value protecting non-tenured members and instead focus on retaining senior teachers. However, if this were the case I would expect to see a negative coefficient for the share of senior teachers dismissed.

[22] Similar concerns extend to the case of endogeneity between the existence of policies that base pay or tenure on student performance and unionization.

Table 3.8: The Effect of Collective Bargaining on Tenure

	SASS			Survey		
Control variables	(4b)	(4c)	(4d)	(5b)	(5c)	(5d)
Demographic characteristics	yes	yes	yes	yes	yes	yes
Census region indicators	no	yes	no	no	yes	no
State indicators	no	no	yes	no	no	yes
	N=8,294	N=8,294	N=8,294	N=282	N=282	$N \leq$282
Dependent variable	Effect of collective bargaining					
No tenure				-0.0756	-0.106*	0.449***
				(0.0509)	(0.0606)	(0.148)
Years to tenure				-0.743***	-0.834***	-0.0166
				(0.260)	(0.265)	(0.138)
Administrator review				0.0382	0.0180	-0.361**
				(0.0435)	(0.0497)	(0.165)
Peer review/observation				-0.0838	-0.115*	0.0853
				(0.0577)	(0.0624)	(0.190)
Teacher portfolio				0.0137	0.0582	0.162
				(0.0669)	(0.0747)	(0.183)
Student performance				-0.163**	-0.120	0.006
				(0.0692)	(0.0847)	(0.182)
Senior teachers dismissed	-0.00394	-0.00924	-0.0195	0.0147*	0.0127	0.003
	(0.00775)	(0.0101)	(0.0129)	(0.0800)	(0.0791)	(0.0277)
Junior teachers dismissed	0.000520	-0.00238	-0.00360	0.00921**	0.00812*	0.00135
	(0.00323)	(0.00539)	(0.0051)	(0.00387)	(0.00420)	(0.00649)

Table reports of β_1 from OLS or marginal effects from probit estimations of equation (3.1) with various dependent and control variables. Robust standard errors are in parentheses. Significance: $*$: 10% $**$: 5% $***$: 1%. Demographic characteristics are total enrollment, racial composition of the student body, the share of students that receive free and reduced lunch, the total number of teachers, racial composition of the teachers, the share of school aged children in the district, the share of college educated residents in the district and per capita income. Columns (4b), (4c) and (4d) use the weights provided with the SASS. Columns (5b), (5c) and (5d) use the survey weights described in the text.

This is particularly the case for the variables that measure log salary at different points in a teacher's career. The IV results for the returns to degree and experience and the non-monetary aspects of the contract are less consistent.

The first set of instrument measures the longevity of state laws granting teachers unionization rights are adapted from Hoxby (1996). States are divided into four groups: (1) states where legislation allowing for union activities was passed prior to 1970; (2) states where such laws were enacted between 1970 and 1980; (3) states with legislation passed between 1980 and 1990; and (4) states where laws do not support teachers' right to unionize by 1990 (the omitted category). The timing of state law may impact the existence of a teachers' union but such legislation is unlikely to be correlated with teacher compensation other than through its impact on unionization (Hoxby, 1996). Additionally, I recreate Hoxby's instruments that measure the existence of legislation that either permits "agency shops" or "union shops."[23] It is important to note, however, that Hoxby's instruments were designed to be used with panel data that spans the period from 1972 to 1992. Using these instruments for my pooled cross section of data from 2003-04 and 2007-08 may be less appropriate. Specifically, using these instruments for my analysis implies that districts in states with longer standing laws and/or more stringent laws are more likely to engage in collective bargaining than districts in states that did not permit unions until the 1980s.

Second, I construct a similar set of instruments from data available for the National Council on Teacher Quality (NCTQ). These measure current state laws regarding teacher unionization rights. Here states are divided into three groups: (1) states where collective bargaining is mandatory; (2) states where collective bargaining is permissible – i.e. neither mandatory nor illegal; (3) states where collective bargaining is explicitly illegal (the omitted category). Additionally, I create instruments that measure whether bargaining over wages is mandatory (some states explicitly list topics that must be part of the collective bargaining process) and whether teachers have the clear right to strike. The benefit of these instruments over Hoxby's is that they may more accurately reflect the laws in 2003-04 and 2007-08. Another benefit is that a two stage least squares (2SLS) estimate based on the current legislation may more accurately reflect the impact of law changes such as those recently enacted in Wisconsin and under consideration in other states. The draw

[23] A union has an agency shop if it collects dues from all teachers in the bargaining unit and a union shop exists if the school district cannot employ teachers who do not become union members.(Hoxby, 1996)

back is that because they are contemporaneous, omitted variables that might impact both teacher pay and laws regarding teachers' unions are more likely to be a problem (that is, the instrument may be as endogenous as the simple indicator for collective bargaining).

I also construct an instrument that measures the percent of workers in all professions in the state who are members of a unions is calculated from the Current Population Survey (CPS). The logic of this instrument is that teachers' unions are more likely to have formed in states where there is a strong culture of unionization. I use data from 1964 because it will not be influenced by contemporaneous events that may impact both teacher compensation and the level of unionization in the state. It is of course still possible that even a lagged measure will be endogenous. Therefore, I construct a final instrument that measures the share of immigrants in the state from 1900-1930 using data from the Integrated Public Use Microsample (IPUMS) data.(Ruggles et al., 2010) The rationale for this instrument is that states with larger shares of immigrants were more likely to pass legislation supportive of unions and more likely to have a strong culture of unionization, however, it is highly unlikely that the share of immigrants in the early 1900s has any impact on current teacher compensation and tenure policies other than through its impact on the existence of teachers' unions.

Overall, the apriori expectation that OLS and probit estimates may be a lower bound of the true impact of unions is supported. There are, however, a few reasons to prefer the simple OLS and probit estimates presented in the main specification. First, the instruments are all state-level variables and thus I am unable to estimate models with state fixed effects. Second, while the instruments generally perform well on diagnostic tests, some doubt is cast on the exogeneity of the instruments since a test of overidentification restrictions based on Sargan's statistic rejects the null hypothesis of non-correlation of the instruments with the errors.

A second empirical issues is that the pooled cross-section ignores the fact that some districts are represented twice. For the outcome variables available in the SASS (but not those available only in my survey), it is possible to create an unbalanced panel that spans two time periods. There are 2,215 districts that are sampled in both 2003-04 and 2007-08. Including a district fixed effect or district random effect can control for unobservable differences between districts. The problem with a district fixed effect is that the only

identifying variation would come from districts that changed unionization status between 2003-04 and 2007-08. Only 235 districts fit this category and, of these, only 74 went from having no union in 2003-04 to having a union in 2007-08, see Appendix Table 7.9.[24] The problem with a district random effect is that, since the data is not designed to be a nationally representative panel, I am unable to use the sampling weight provided by the SASS. Results analogous to Table 3.4 that include a district random effect – but ignore the weights provided by the SASS – are presented in Appendix Table 7.10. These results show that the district random effect makes almost no difference so I prefer the results presented in the main text that preserve the information contained in the sampling weights.

3.5 Discussion

This study offers new evidence on the impact of teachers' unions on compensation and tenure policy. I use data from the 2003-04 and 2007-08 SASS as well as new data from a survey that I conduct and find that contracts negotiated via collective bargaining differ from contracts in non-union districts, but perhaps in some different ways than union opponents might expect. Much of what I find is consistent with economic theory and previous empirical work, however, I offer new detail that will be of interest to policy makers who are considering legislation to limit collective bargaining rights as well as those who are considering reforms to teacher compensation and tenure policies.

Theory predicts that unions will favor policies that benefit the median teacher. I find that a teacher with ten years of experience (likely close to the median teacher), earns a higher salary in unionized districts both when the union has collective bargaining rights and when the union has only meet and confer privileges. The difference is larger, however, where teachers have collective bargaining rights largely because of greater returns to experience, i.e. step increases. In districts with collective bargaining, step increases are approximately 26% higher than in districts without collective bargaining. This appears to be the main contract difference attributable to collective bargaining. The greater returns to experience yield salaries for teachers with ten years of experience that are approximately 5% higher than in districts with only meet and confer privileges. This magnitude is consistent with

[24] Although this is a small sample, future research should explore the impact of a change in collective bargaining status on the various contract measures studied in this paper.

Hoxby (1996).

Hoxby (1996) argues that the union wage premium is the result of union rent-seeking rather than increased productivity. If additional experience is not associated with increased productivity, then my finding supports this argument. The literature on this point is somewhat mixed. In general, additional experience does not lead to increased productivity. However, in the first three years, teachers do appear to make large gains in their ability to increase student achievement. Some research suggests that the gains to experience level off but remain positive through year ten (Hanushek, 2003). One weakness of the current study is that I do not know the exact wage-experience profile. Instead, I average the returns to experience equally over the first ten years. If it is the case that unionized districts have steeper wage-experience profiles in the first three years and then the profile levels off, this may be consistent with what we know about gains in teacher productivity over time.[25]

Also, higher returns to experience may be rational if delaying wage increases provides an incentive for teachers to remain in the profession longer and thus decreases turn-over rates. To be effective, this would have to be paired with dismissal of ineffective teachers since providing an incentive for ineffective teachers to stay is counter productive. While I find some evidence that unionized districts dismiss slightly more teachers for poor performance, this result is preliminary. I find that, in general, tenure policies are very similar in union and non-union districts. The notable exception is that districts with collective bargaining offer tenure protections almost one year sooner than their non-union counterparts. Presumably this gives districts less time to learn about a teacher's productivity and increases the likelihood that they will tenure an ineffective teacher. This is a topic ripe for further research.

Turning to P4P, theory and evidence from the private sector indicate that unions will oppose output based P4P but be more open to input based P4P. I offer some of the first evidence from the public sector on this topic. I find that where teachers are unionized, rewards for measures of teacher inputs, be that in the form of a masters degree other credentials or professional development, are larger. This is consistent with results in West and Mykerezi (2011). Union opposition to output based P4P is most evident when examining the use of rewards for administrator review. Administrator review is difficult to classify

[25] Vigdor (2008) advocates for a very steep salary schedule that rewards experience in the first few years but then levels off.

as purely out based P4P since administrators likely consider a mix of teacher actions and student outcomes. It is input based in so much as it measures teacher effort. It is output based in so much as it is not directly in the teacher's control. This lack of control will make risk-averse agents wary of basing compensation on these reviews. If the median teacher is risk-averse, unions will oppose rewards based on administrator review unless they are sufficient large in expectation. Anecdotally, teachers' unions are often very concerned that uninformed or even vindictive administrators will not be able to conduct fair evaluations.

The fact that unions oppose rewards based on administrator review will be of interest to policy makers since there has been a recent push to increase the use of this type of P4P. For instance, Taylor and Tyler (2011) report on a P4P program in Cincinnati and Sojourner et al. (2012) have work forthcoming on a P4P program in Minnesota both of which include rewards for administrator review. Both of these programs were negotiated in close cooperation with the local teachers' unions so studying how they differ from other programs may be of particular interest.

In sum, this study adds important detail to our understanding of teacher contracts. I offer new data with more granular measures of P4P and tenure policies than has been previously available. It is worth noting that the magnitudes that I find for the union impact on the specifics of P4P and tenure reform are not overwhelming. Teacher contracts in non-union districts are nearly identical to those where teachers are organized, either with or without legal collective bargaining rights. The dominant contract in education everywhere bases compensation on a single salary schedule and awards tenure after three years based on administrator review. Deviations from this norm are rarer in unionized districts but they are very rare events regardless of unionization status. For instance, districts with collective bargaining are only on the order of 6% less likely to link compensation to administrator review than are non-union districts. Further, district with only meet and confer privileges are around 3.5% less likely to link compensation to administrator review than are non-union districts. This suggests that collective bargaining, while important, is far from the only impediment to reform. Opposition from teachers (even when not voiced via formal collective bargaining), a wait-and-see attitude on the behalf or administrators and even simple inertia are likely also obstacles to reform.

Chapter 4

Are teachers overpaid or overworked? New measures of market hours.

4.1 Introduction

Influential studies have drawn attention to teacher quality (Nye et al., 2004; Rivkin et al., 2005) and policy makers have taken note. The Race to the Top Fund provides in excess of $4 billion dollars in part to aid in efforts to "recruit, reward and retain quality teachers" (Race to the Top Program Description). The efficacy of policies aimed at improving teacher quality depends on a number of important teacher labor supply decisions. Who becomes a teacher, how much and how hard they work, and whether they stay in the profession are all issues that require an understanding of teachers as labor market participants. Despite significant amounts of research and money devoted to improving teacher quality, there remain some basic deficits in our knowledge about teacher labor markets. One notable deficit is our lack of clarity on teachers' wages and how they compare to wages in other occupations.

This is frustrating to policy makers and voters, consider the following excerpt from a letter to the editor in the Minneapolis Star Tribune (2011):

> "Recent articles about a proposed pay freeze for teachers have contained predictable posturing from the usual suspects - politicians, union leaders, teachers and school administrators - but little information about current teacher pay. [...] How many hours per year do teachers work for this pay? How many unemployed teachers are seeking positions at current compensation levels? And how does pay in the public sector compare to pay in the private sector? [...] Let's put the relevant facts on the table and let the people decide whether a freeze is unfair or overdue."

Unfortunately, it is not only the popular press that presents unclear and conflicting information about teacher pay. The academic literature on the topic is not much better. At the core of the problem is the fact that the research community has good data on teachers' annual salaries but only a very hazy understand of teachers' hours of work. As this frustrated letter writer alludes to, this leads to conflicting estimates of teachers' hourly wages and hence very different policy prescriptions. Basic supply and demand analysis shows that if wages are too low, there will be a shortage of high quality teachers. This motivates many to call for increased wages for teachers (Allegretto et al., 2004; Temin, 2003). Others counter that teacher wages are already high in comparison to similarly educated workers and argue that raising wages will only produce a glut of low quality candidates (Richwine and Biggs, 2011; Podgurksy, 2003; Ballou and Podgursky, 2002).

Within this debate there are a number of arguments about whether and how to adjust teachers' wages to account for the length of the school day and year. These arguments often hinge on different assumptions about how much time teachers spend on work related activities outside of school hours (Nelson and Podgursky, 2003; Podgursky and Mishel, 2005). At one extreme, economists use administrative data on contract hours and assume that teachers do not work at all beyond what is minimally required (Podgursky and Tongrut, 2006).[1] More commonly, they use self reported data from surveys such as the Current Population Survey (CPS) and the Schools and Staffing Survey (SASS) which ask respondents about a usual or typical work week (Flyer and Rosen, 1997; Loeb and Page, 2000; Allegretto et al., 2008).

[1] Contract hours are the hours that a teacher is formally required to work as opposed to informal expectations about hours of work.

In this study, I offer new measures of hours of work from the American Time Use Survey (ATUS). The time diaries collected by the ATUS are a more reliable way to estimate hours of work than either contract data or surveys such as the CPS or the SASS.[2] The ATUS provides a unique opportunity to investigate the time teachers spend working and whether teachers more likely to over report their hours in the CPS than other workers. My analysis of the time diary data from the ATUS has the potential to close the debate over whether teachers work more or less than the average worker and allow for a more accurate comparison of teachers' wages to wages in other sectors.

I find that teachers work an average of 34.5 hours per week annually. During the school year they work an average of 38.0 hours per week and during the summer they work an average of 21.5 hours per week. Teachers work more than they are required to work by contract, but less than self reported usual or typical hours. I find that teachers are more likely to overreport their usual hours of work in the CPS than workers in other occupations and conclude that this is likely because of an uneven work year. Finally, I use time diary data from the ATUS to compare teachers' wages to wages in other sectors. I find that high school teachers earn 11% less than full time workers with at least a Bachelors degree in other occupations but elementary, middle and special education teachers are not underpaid relative to full time workers with at least a Bachelors degree in other occupations. I discuss results by sector, gender and separately for workers with Bachelors degrees and Masters degrees.

My findings have implications beyond the debate over teacher pay. This study suggests that the measures of market hours in the CPS may be systematically biased. Teachers are not the only workers with uneven work years, and a failure to account for this may lead economists to faulty calculations of hourly wages for other occupations as well. Even if the problem is limited only to teachers, teachers comprise a large segment of the labor force – accounting for over 18% of college educated women who work full time.[3] Biased measures of hours of work therefore may impact calculations of wage gaps across levels of education,

[2] Juster and Stafford (1991) write (page 473) that "The methodology for collecting time allocation data has been well developed at this point, and the main characteristics of optimum methodology are not in dispute. The only way in which reliable data on time allocation have been obtained is by use of time diaries, administered to a sample of individuals in a population and organized in such as way as to provide a probability sample of all types of days and of the different seasons of the year." The ATUS fits this optimal methodology, the CPS and the SASS do not.

[3] Authors calculations based on the final sample described in the data section of this paper.

gender and race.

The chapter proceeds as follows. In the next section I briefly describe similar work by others. Section 4.3 describes the data. Section 4.4 describes my methodology for constructing time diary measures of hours of work per week by occupation and presents results that compare teachers hours of work and propensity to over report hours in the CPS to other occupations. Section 4.5 shows the impact of using time diary measures of hours of work to compute the wage gap between teachers and other occupations as well as the wage gaps by level of education, gender and race. Section 4.6 concludes with a discussion of the policy implications and suggestions for future research.

4.2 Previous Literature

Other than the ATUS, the only other time diary data on teachers was collected and analysed by Drago et al. (1999). They find that the average elementary school teacher works 9.7 hours per day, almost two hours more than what is required by contract. Drago et al. (1999) use data from a survey that is much smaller than the ATUS and does not include non-teachers so it is not possible to compare teachers to other workers. They were also unable to compare diary measures to self reported usual or typical hours of work.

I am aware of only one other paper that uses ATUS data to examine teachers' and their work patterns. Krantz-Kent (2008) provides a short "visual essay" which summarizes teacher work patterns using ATUS data. She finds that teachers' hours of work vary throughout the year. Not surprisingly, teachers are less likely to work during the summer months than they are during the school year. Almost half of all teachers, however, report some work during the prior week for interviews conducted in July and over 70 percent report some work during the prior week for interviews conducted in June and August. Krantz-Kent (2008) finds that teachers are more likely than others to work at home and to work on Sundays, but teachers work fewer hours during the week and on Saturdays. The net result is that teachers spend, on average, for all days of the week, 18 fewer minutes per day working than other professionals.[4] The analysis stops at summary statistics and

[4]The definition of "other professionals" in Krantz-Kent (2008) includes health care professionals, business and finance operations professionals, architects and engineers, community and social service professionals, managers and unspecified "others." This is a bit ambiguous, for instance, it is not clear how she categorizes

does not consider the implications for wages nor compare the hours of work reported in the ATUS to the CPS.

I am not the first researcher to use the ATUS to estimate over reporting in the CPS; however, there is no other study that looks at over reporting by occupation. Juster and Stafford (1991) note that surveys like the CPS that ask about usual weeks of work are likely to have valid responses only when daily work patterns have regular schedules. Teachers do not have regular work patterns for a number of reasons, most notably because a teacher's contract generally requires nine or ten months of work rather than twelve. When asked to recollect a usual day it is unlikely that teachers will average in their time off during the summer months. This makes teachers more likely than others to overestimate their hours of work. Allegretto et al. (2004) cite personal correspondence with the Bureau of Labor Studies (BLS) that urges caution when attempting to interpret usual hours of work data for teachers (they mention flight crews, sales representatives and truck drivers as other occupations that likely suffer from the same estimation challenge).

The study most similar to my analysis of over reporting and its impact on estimating hourly wages is Frazis and Stewart (2004).[5] They compare the "usual hours of work" variable in the CPS to diary data in the ATUS (without regard for occupation) and find that all workers over report by an average of three hours per week. Frazis and Stewart find that the "actual hours of work" variable in the CPS is much closer to diary data in the ATUS. Here the over reporting is closer to one hour per week. They also find that women and more educated respondents are more likely to over report hours in the CPS relative to diary in the ATUS. They conclude that accounting for over reporting increases the college-high school earnings ratio by 4.1% and the female-male hourly earnings ratio by 5.4%.[6]

---

professors.

[5]Other discussions of over reporting that rely on different data sources include Baum-Snow and Neal (2009) who compare hours of work for part time workers in the CPS and the American Community Survey and Robinson and Bostrom (1994) who compare hours of work in the CPS to earlier time diary studies conducted by the University of Michigan and the University of Maryland. Podgursky and Tongrut (2006) focus on over reporting by teachers by comparing the CPS and the National Compensation Survey.

[6]Although it is more common to talk of the male-female ratio, Frazis and Stewart (2004) state their finding in terms of the female-male ratio. Presumably by "increase" they mean that the wage gap widens, i.e. the female-male ratio becomes more negative.

4.3 Data

The data for this paper come from the Current Population Survey (CPS) and the American Time Use Survey (ATUS) extracted via the ATUS-X(Abraham et al., 2011). The CPS is a probability sample of 60,000 households conducted monthly by the Census Bureau for the Bureau of Labor Statistics. The ATUS is a diary survey that has been collected for a subsample of individuals included in the CPS since 2003. Specifically, 1/8 of the households selected by the CPS retire permanently from the CPS sample each month and these households become eligible for the ATUS two months later. Respondents are offered $40 for participating and the response rate is approximately 52 percent.[7]

Only one person in each household is surveyed and they are only asked about their activities for the prior day (4 a.m. yesterday to 4 a.m. today). Respondents are asked what they did and when, along with who they were with and where they were at the time. They can only indicate one activity at a time. While time use information is collected only for the survey respondent, CPS data are available for all members of the household.

The ATUS provides survey weights and replicate weights which I use extensively in my analysis. The weights compensate for three important aspects of the data collection process: (1) The ATUS is a stratified random sample that oversamples some demographic groups, (2) the ATUS sample is not uniformly distributed across days of the week, specifically weekends are oversampled, and (3) the response rates differ across demographic groups and days of the week.r:A (2012).

I pool ATUS data from 2003 through 2010. I limit the sample[8] to include only respondents that report being employed at the time of his or her ATUS interview because the variable of interest, whether the respondent is a teacher, is available only for this group. A respondent is designated as a teacher if his or her primary job is coded as either an elementary, middle, high school or special education teacher.[9] I further limit the sample to

[7]Research into non-response bias does not suggest any particular problems for this study (Abraham et al., 2006). Additionally, I check non-response by occupation and find that teachers have a relatively high response rate of 66 percent.

[8]To be exact, I do not limit the sample in the sense of dropping observations because this would impact the standard error calculations using the replicate weights provided. Instead I use the *svy, subpop* option in STATA to specify the sample I describe in this paragraph.

[9]Preschool, kindergarten, postsecondary and "other" teachers and instructors are not included as teachers.

include only full time workers because time use patterns for full and part time workers are very different and comparisons across occupations would be problematic if more (or fewer) teachers work part time than do workers in other occupations. The CPS and the ATUS do not report weekly earnings for persons who are self-employed so these observations (along with any observation missing weekly earnings for any other reason and those with zero weekly earnings) are also dropped from the final sample. Lastly, I include only workers with at least a Bachelors degree. I do this because teaching requires a four year degree and limiting the sample to college educated workers provides a better comparison group when estimating the wage gap between teachers and other occupations. This decision assumes that people do not decide between not attending college and attending college to pursue teaching but rather that they decide to attend college and subsequently choose an occupation.[10]

There are 2,129 teachers in the final sample and 16,025 non-teachers (teachers comprise 11.7% of the observations). Table 4.1 shows summary statistics for teachers and non-teachers. Teachers are slightly more white than non-teachers. More notably, teachers are much more likely to be female than non-teachers and they are twice as likely to have a Masters degree than non-teachers, although they are less likely to have a PhD or professional degree. On average, teachers earn $350 (USD 2010) per week less than other full time workers with at least a Bachelors degree. This yields a naive estimate of the "teacher wage gap" of 24%, i.e. teachers earn 24% less than other college educated workers ($350/$1,426). Of course, this does not control for demographic characteristics and, as just noted, teachers are different from other college educated workers in a number of ways. In the results section I present wage regressions that control for the demographic variables listed in Table 4.1.

Table 4.2 shows the breakdown of teachers by sector and assignment. Of the 2,129 teachers in the final sample, 1,841 or 86.4% work in the public sector. There are roughly twice as many elementary/middle school teachers as their are high school teachers. Special

[10]From a policy perspective, this is in keeping with discussions about how to best recruit high aptitude college students to teaching. Programs like Teach for America or The New Teacher Project aim to take top college graduates and entice them to join the teaching ranks. It is generally assumed that quality teachers are currently not in the teaching pool because they are engaged in other occupations that require a Bachelors degree, not because they lack a Bachelors degree.

Table 4.1: ATUS Summary Statistics

Variable	Teachers $N = 2,129$ Mean (SD)	Non-Teachers $N = 16,025$ Mean (SD)	Difference
Age	42.2 (10.7)	42.3(10.5)	0.1
Female	0.77	0.46	0.32***
White	0.89	0.82	0.07***
Masters degree	0.48	0.24	0.24***
PhD degree	0.01	0.06	-0.05***
Professional degree	0.02	0.05	-0.03***
Weekly earnings	$1,076 (482)	$1,426 (759)	-350***

Significance *** 1%, ** 5%, * 10%. Sample includes ATUS respondents with at least a Bachelors who are full time workers with positive weekly earnings. Teachers include elementary, middle, high school and special education teachers. Weekly earnings are in 2010 dollars.

Table 4.2: Teachers by Sector and Assignment

	Assignment			
Sector	Elementary/middle	Secondary	Special education	Total
Public	1,168	503	170	1,841
Private	170	97	21	288
Total	1,338	600	191	2,129

education teachers comprise a small share of the sample.

When studying teachers, one must be particularly concerned about the summer months. Data for people who are employed but currently absent from work for reasons such as vacation days, illness, and maternity leave are included in the final sample. Teachers who are employed but not at work during summer months should be counted as employed and on vacation. It is possible that during the summer some teachers report being unemployed (and thus are not in the final sample) or that they are categorized into another occupation during the summer months (and thus are in the final sample but categorized as a non-teacher), e.g. a teacher who is a waitress during the summer will be categorized as a waitress rather than a teacher.

Figure 4.1 shows the share of ATUS respondents who are teachers by month. Figure 4.2 shows the weighted share of teachers by month. This should approximate the share of teachers in the total population. The fact that the share of teachers falls during the summer (particularly in July) could be evidence that teachers are missing or miscategorized during the summer. Alternatively, it may be that there are fewer teachers during the summer because all job transitions in education happen in the summer. Unlike other occupations where retirements, layoffs and hiring take place across the calendar year, teachers generally retire, are laid off and hired only when classes are not in session. If retirements and layoffs happen early in the summer and hiring takes place later in the summer, this would create a drop in the share of teachers in July. I assume that each of these explanations contributes to the drop in the share of teachers in the summer to some extent.[11] In any case, it seems safe to assume that teachers who are working a different job during the summer or are laid off and hoping to be rehired will not devote any more hours to teaching than teachers who are continuously categorized as such. Thus, my calculations of the average hours devoted to teaching during the summer months will be an upper bound of the true number of hours devoted to teaching during the summer months.

4.4 Results

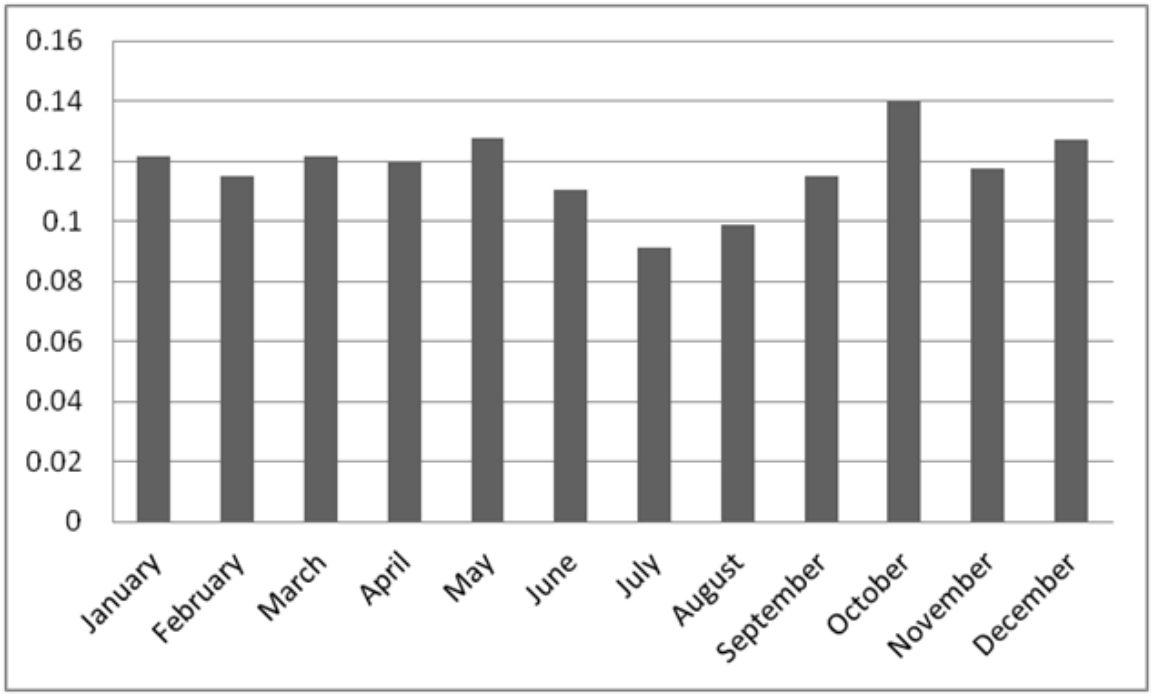
4.4.1 Average Hours of Work by Occupation

Since the ATUS collects diary data for a single day, it is impossible to calculate the hours worked over an entire week for any one individual. Instead, I calculate a weighted average hours of market work on weekdays for all respondents within an occupational category to create synthetic average work weeks by occupation.[12]

[11] I attempt to estimate the relative importance of these explanations using teacher turnover data from other sources. The average number of teachers in the ATUS drops 23% in the summer. Keigher (2010) uses data from the SASS and finds that 9% of teachers exited the profession in 2008. Harris and Adams (2007) find a similar turnover rate using CPS data. Additionally, Keigher (2010) reports that 7% of teachers switched jobs within teaching, therefore, approximately 16% of teachers are potentially either leave or have a gap in employment and may no longer be counted as teachers during the summer due to normal turnover and job transition rates. This suggests that normal turnover rates account for as much as two thirds of the dip observed in Figure 4.1.

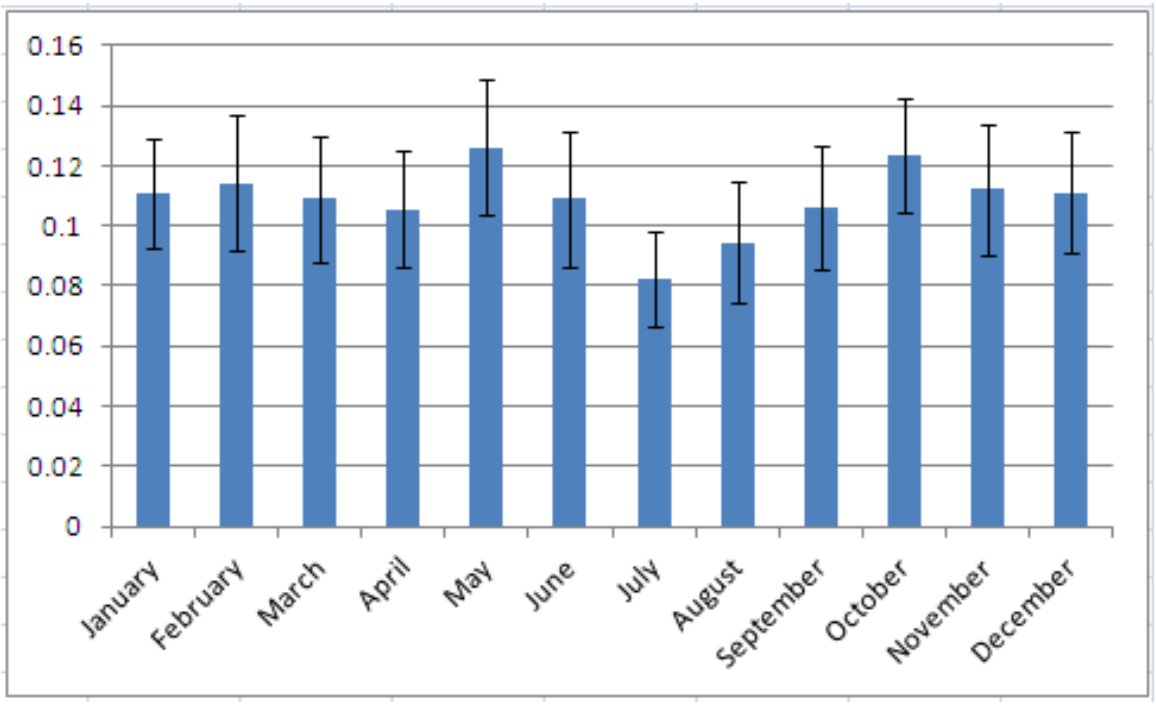
[12] This methodology is similar to Robinson and Bostrom (1994) and others who use time diaries to compare time use by demographic characteristics.

Figure 4.1: Share of ATUS Respondents Who are Teachers, by Month



Each bar shows the share of the final sample ATUS respondents who are teachers by the month of the respondent's ATUS interview. The final sample includes all ATUS respondents with at least a Bachelors degree who are full time workers and have positive weekly earnings. Teachers include elementary, middle, high school and special education teachers.

Figure 4.2: Share of ATUS Respondents Who are Teachers, by Month – Weighted



Each bar shows the weighted share of respondents in the ATUS who are teachers by the month of the respondent's ATUS interview. The weights should approximate the share of the population who are teachers.

I call this measure "diary hours of work" to distinguish it from the "usual hours of work" or "hours of work last week" variables reported in the CPS. I define weekly diary hours of work, $DHrs_o$, for occupation o as:

$$DHrs_o = \frac{\sum_{i=1}^{N}[(Dhrs_{io} * 7)(w_i)]}{N} \tag{4.1}$$

where $Dhrs_{io}$ is hours of work on the respondent's main job reported in the time diary for the N respondents in occupation o.[13] Weights for each respondent provided by the ATUS, w_i, are used so that $DHrs_o$ is adjusted for the sampling scheme. Most importantly, they adjust for the fact that weekends are oversampled. Standard errors and confidence intervals are calculated using replicate weights provided by the ATUS.

Not surprisingly, teachers' diary hours of work per week vary across the calendar year. Figure 4.3 shows diary hours of work calculated using equation (4.1) for teachers and non teachers for each calendar month. Here we can see the expected dip in hours of work over the summer. The average for teachers during the summer months (June, July and August) is 21.5 hours/week ($SE = 1.7$) and the average during the school year is 38.0 hours/week ($SE = 0.8$).[14] The average over the entire calendar year for teachers is 34.5 ($SE = 0.7$) hours/week. The average over the entire calendar year for non-teachers is 39.9 hours/week ($SE = 0.3$).[15]

4.4.2 Over Reporting

Recall that all ATUS respondents are also CPS respondents and CPS respondents are asked about both their usual hours of work per week and their hours of work last week.[16] ATUS

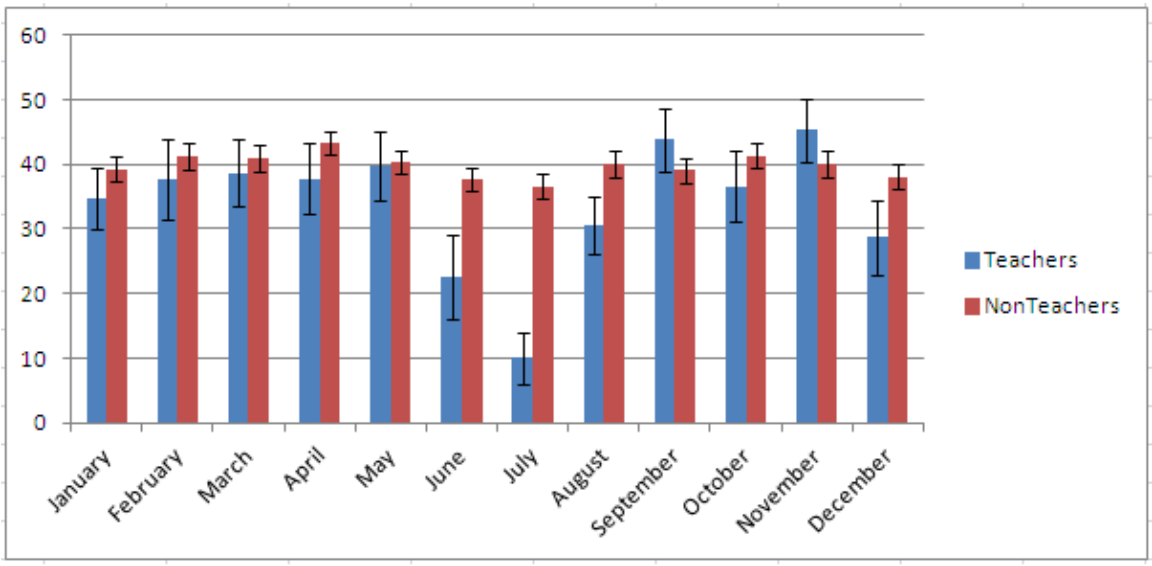
[13] To be exact, the ATUS reports hours of work in minutes but I divide by 60 and present results in hours.

[14] The school year generally ends in mid-June and begins in mid-August so the 21.5 hours/week includes an average of some full work weeks and some vacation weeks. Also, some teachers work summer school and almost all teachers spend at least some time engaged in planning and professional development activities during the summer months. Recall that the estimate of summer hours is likely an upper bound since there may be teachers who are missing or miscategorized during the summer and I assume these teachers spend no more time on activities related to teaching than their peers who are continuously categorized as teachers.

[15] Non teachers also work slightly fewer hours in the summer months. The average across all non teachers is 40.4 hours/week ($SE = 0.3$) during the school year and 38.1 hours/week ($SE = 0.6$) during the summer.

[16] These questions were revised in 1994. Prior to 1994, respondents were only asked about their work last week and the CPS attempted to correct for time off and overtime to figure out a usual work week.

Figure 4.3: Weekly Diary Hours of Work, by Month



Each bar shows the diary hours of work per week (on the respondent's main job) for teachers and non teachers by the month of their ATUS interview. Diary hours are calculated using equation (4.1). Confidence intervals are calculated using the replicate weights provided by the ATUS. The final sample includes all ATUS respondents with at least a Bachelors degree who are full time workers and have positive weekly earnings. Teachers include elementary, middle, high school and special education teachers. Non teachers are all other occupations.

respondents are also asked about their usual hours of work again at the time of their ATUS interview. The modal response for questions about usual hours and hours of work last week is 40 hours per week for both teachers and non-teachers. Figure 4.4 shows the distribution of usual hours and hours of work last week for teachers and non-teachers.[17] The most noteworthy point made by this figure is how incredibly similar teachers and non-teachers are in their self-reported hours of work.

For usual hours of work, respondents can report that their hours vary.[18] Interestingly, despite the fact that teachers' work weeks look very different in the summer than during the school year, teachers are *less* likely than others to report that their hours vary, 4% of teachers report that their hours vary compared to 5% of non-teachers.[19]

I define over reporting OR_o in occupation o as:

$$OR_o = \frac{\sum_{i=1}^{N}[Uhrs_{io} - (Dhrs_{io} * 7)(w_i)]}{N} \tag{4.2}$$

where $Uhrs_{io}$ is the usual hours of work reported by the N respondents in occupation o. As before, $Dhrs_{io}$ is the diary hours for each respondent and w_i are weights provided by the ATUS.

It is more interesting to look at over reporting as a percent of total hours so next I modify the definition of over reporting to:

$$ORpercent_o = \frac{\sum_{i=1}^{N}[\frac{Uhrs_{io}-(Dhrs_{io}*7)}{Uhrs_{io}}(w_i)]}{N} * 100 \tag{4.3}$$

Figure 4.5 shows $ORpercent_o$ by occupation. The occupations shown are the most common in the final sample, i.e. they are common occupations for full time workers with

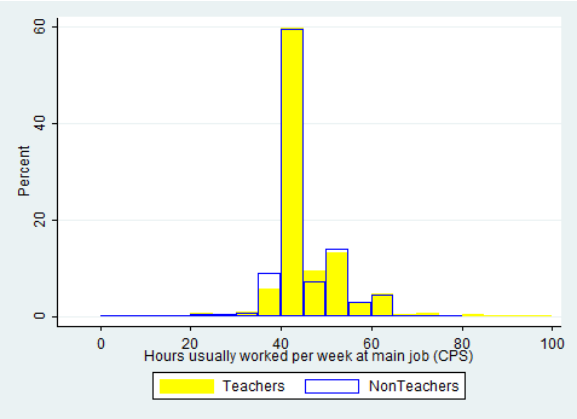
[17] For this figure I use the responses from the final CPS interview because the actual hours of work question is not asked as part of the ATUS. A few respondents switch between teacher and non-teacher but otherwise the sample is identical to the one described in the data section.

[18] CPS documentation indicates that if the respondent asks for a definition of "usual," interviewers are instructed to define the term as more than half the weeks worked during the past four or five months.

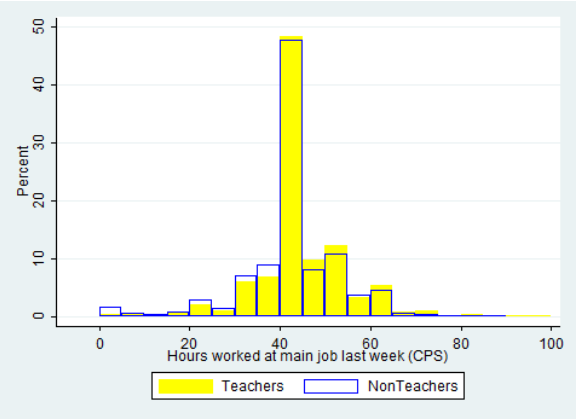
[19] This difference is statistically significant at the 1% level. When asked about their usual hours of work per week in the ATUS, only 2.5% of teachers report that their hours vary and only 2.7% of non-teachers report that their hours vary. This difference is not statistically significant.

Figure 4.4: Self Reported Hours of Work, Teachers and Non-Teachers

Usual hours of work



Actual hours of work



This figure shows the distribution of responses to the questions about respondents' usual hours of work per week and hours of work last week. Data for both are from the final CPS interview. The sample includes all ATUS respondents with at least a Bachelors degree who are full time workers and have positive weekly earnings. Additionally, respondents who changed occupations between the time of their final CPS interview and their ATUS interview are excluded. Teachers include elementary, middle, high school and special education teachers. Non-teachers include all other occupations.

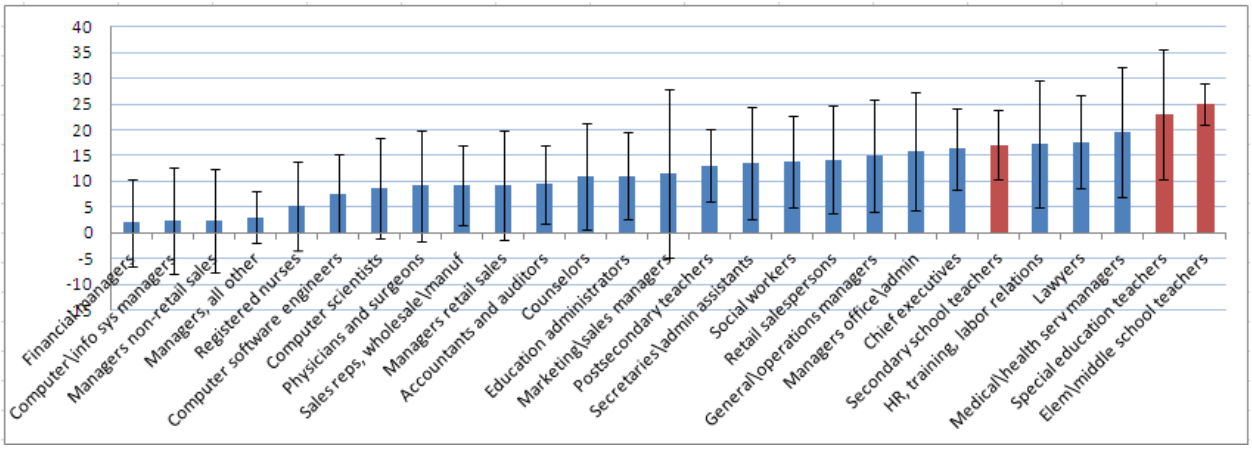
at least a Bachelors degree.[20] The three categories of teachers (elementary/middle school,[21] are shown in a darker color. Over reporting is quite common; however, teachers over report their usual hours of work more dramatically. Secondary, elementary/middle and special education teachers over report their usual hours of work in the ATUS by 17% ($SE = 3$), 25% ($SE = 2$) and 23% ($SE = 6$) respectively. The average across all teachers is 22% ($SE = 1.6$), the average across all other occupations is 10% ($SE = 0.6$).

One might hope that the CPS question which asks about hours of work last week rather than a usual week is akin to asking about hours of work in a time diary survey. Unfortunately, the fact that the time diary is separated from the CPS interview by at least two months makes it impossible to judge this for individual respondents. Using occupational averages, I find that over reporting remains a problem even when asking

[20] Each occupation has at least 190 respondents in the final sample.

[21] The decision to group middle school teachers with elementary school teachers is entirely based on the occupation categories provided by the CPS and ATUS. This is also why Kindergarten teachers are excluded from the category of elementary school teachers.

Figure 4.5: Over Reporting of Usual Hours of Work, by Occupation



Each bar shows over reporting, $ORpercent$, from equation (4.3), confidence intervals are calculated using replicate weights provided by the ATUS. The occupations included are the most common in the final sample.

about hours of work last week.[22] Over reporting is less severe than when asking about usual hours of work, however, significant discrepancies remain. Most importantly for the current study, elementary/middle and special education teachers remain the most likely to over report. Elementary/middle school and special education teachers average hours of work last week differ from their time diary hours of work by 16% and 19% respectively. Secondary school teachers over report their actual hours of work last week by 13%. The average over report of hours of work last week across all teachers is 15.7%, the average across all other occupations is 6.5%.

When expressed as hours rather than percentages, the average over reporting across all full time workers with at least a Bachelors degree is 5.4 hours per week when comparing usual hours of work in the ATUS to time diary hours and 3.0 hours per week when comparing hours of work last week in the CPS to time dairy hours. These magnitudes are

[22]Specifically, I subtract the average of $Dhrs_{io}$ for all N respondents in occupation o from the average hours of work last week for all N respondents in occupation o. The draw back to this approach is that, while it is intuitively quite similar to the approach described above, it does not calculate an over reporting figure for each respondent so the weights provided by the ATUS are no longer usable. Instead, I crudely weight to account for the fact that weekdays comprise half of the sample and weekends comprise half of the sample. This ignores any other sampling variability and replicate weights so the results are not as exact.

larger than those found by Frazis and Stewart (2004). They find that among all workers, the average over reporting for usual hours of work is three hours per week and the average over reporting for hours of work last week is one hour per week. However, they also report that education is positively related to over reporting which may explain why my sample of only college educated workers shows more pronounced over reporting.[23]

In sum, teachers are more likely than others to over report their usual hours of work and hours of work last week. It is unlikely that teachers are, on average, more dishonest or forgetful than other workers. Instead, I suspect that teachers have a more difficult time answering questions about their usual hours of work because their work weeks vary across the calendar year. For instance, if a teacher is asked in November about her usual hours of work, she likely thinks about her usual hours of work for the last month or so rather than rolling her summer hours into the calculation. As for hours of work last week, the CPS practice of choosing a reference week to avoid holidays may make it more likely that teachers are interviewed for the CPS after a full week of work while their ATUS interview falls on a week with time off. If teachers have more time off than other workers, this would bias the hours of work last week variable without any malfeasance on the part of teachers.

4.5 Implications for Wage Calculations

The above analysis has important implications for researchers and policy makers interested in wage comparisons. Consider a researcher using CPS data to compare wages across occupations. The CPS reports weekly earnings, $Earn_i$ for each worker i, and usual or actual hours of work per week. Most researchers will use the usual hours of work variable, $UHrs_i$ to calculate an hourly wage, $Wage_i$, according to the simple formula:

$$Wage_i = \frac{Earn_i}{UHrs_i} \tag{4.4}$$

[23] In support of this assertion, when I calculate over reporting for all workers with at least a high school diploma, I find that usual hours are 4.4 hours per week larger than diary hours and hours of work last week are 2.1 hours per week larger than diary hours. This sample is still more educated than the full population since it excludes workers with less than a high school diploma and part time workers. It is likely that the fact that education is positively associated with over reporting is driven by the fact that more education is positively associated with being a salaried employee rather than an hourly employee and salaried workers are more likely to over report than hourly workers.

I have shown that over reporting of usual hours of work is systematically biased by occupation and that the usual hours of work variable is particularly problematic for workers, such as teachers, whose hours vary throughout the calendar year. However, if the weekly earnings data for teachers (and other workers with uneven schedules) already accounts for the variable hours, equation (4.4) will yield an accurate hourly wage. For example, consider a hypothetical worker who works 40 hours/week *every other week* for 52 weeks and earns $52,000. If he reports that he usually works 40 hours/week and earns $1,000 weekly his hourly wage will be, $25/hour, half what it should be. If, on the other hand, he reports that he usually works 40 hours but adjusts his weekly earnings to $500, then the hourly wage will accurately reflect $50/hour. If teachers do the mental arithmetic needed to adjust their weekly earnings to account for variable hours, or if the CPS makes this adjustment to the data after the interview, then the over reporting of usual hours of work is of little concern to labor economists interested in calculating an hourly wage. I show, however, that the weekly earnings data does not properly account for teachers' variable work hours and thus equation (4.4) is measured with error.

CPS respondents are asked to report earnings in the time period they prefer for example, hourly, weekly, biweekly, monthly, or annually. Over 70% of teachers elect to report annual earnings. Allowing respondents to report in a periodicity with which they were most comfortable was added to the CPS in 1994. This improved on the previous procedure which only gave respondents the option to report hourly wages or weekly earnings and was introduced to give the CPS a better chance at calculating accurate weekly earnings since they do not rely on respondents to do the necessary mental arithmetic. Despite this improvement, I show that problems remain.

Respondents who elect to report annual earnings are then asked how many weeks they worked for this salary. Using this information, the CPS converts the annual salaries into weekly earnings.[24] It is unclear, however, whether this refers to weekly earnings for the school year or weekly earnings for the entire calendar year. That is, are teachers and/or the CPS accounting for summers off? I find evidence that they are not.

Figure 4.6 shows the distribution of responses for the number of weeks worked for

[24]Looking at the final earnings per week variable reported in the CPS, I find that teachers who report weekly earnings average $1,036.3 (USD 2010) per week and teachers who report annual salaries average $1,131.6 (USD 2010) per week (these are statistically identical, $Pr(T > t) = 0.9180$.

teachers and non-teachers who report annual earnings.[25] The modal response for both teachers and non-teachers is 52 weeks. This indicates that most teachers view summers as time working. A minority of teachers report between 36 and 44 weeks. These responses more accurately reflect the school year. One reason most teachers report working 52 weeks may be that most teachers are given the option to receive their pay spread over the school year or the calendar year and anecdotal evidence suggests that the majority choose the latter.[26] I suspect that most teachers are reporting weeks paid rather than weeks worked. A simple back of the envelope calculation supports this assertion. The average salary for teachers in 2009-10 was $55,350 (National Center for Education Statistics) which is 53 weeks at $1,036.6 per week and 49 weeks at $1,131.6 per week.

Since the vast majority of teachers report working 52 weeks, if researchers calculate a wage using equation (4.4), they are implicitly (and erroneously) assuming that hours per week are consistent across all weeks during the calendar year. Instead of using equation (4.4) to calculate hourly wages, I use:

$$Wage_i = \frac{Earn_i}{DHrs_o} \tag{4.5}$$

where $Earn_i$ is the CPS weekly earnings variable adjusted to account for inflation and $DHrs_o$ is the time diary measures of hours of work described in equation (4.1) for occupation o. This replaces the individual's estimate of his or her usual hours of work with the estimate I obtain using the ATUS time diaries for all workers within a given occupation.

I also use:

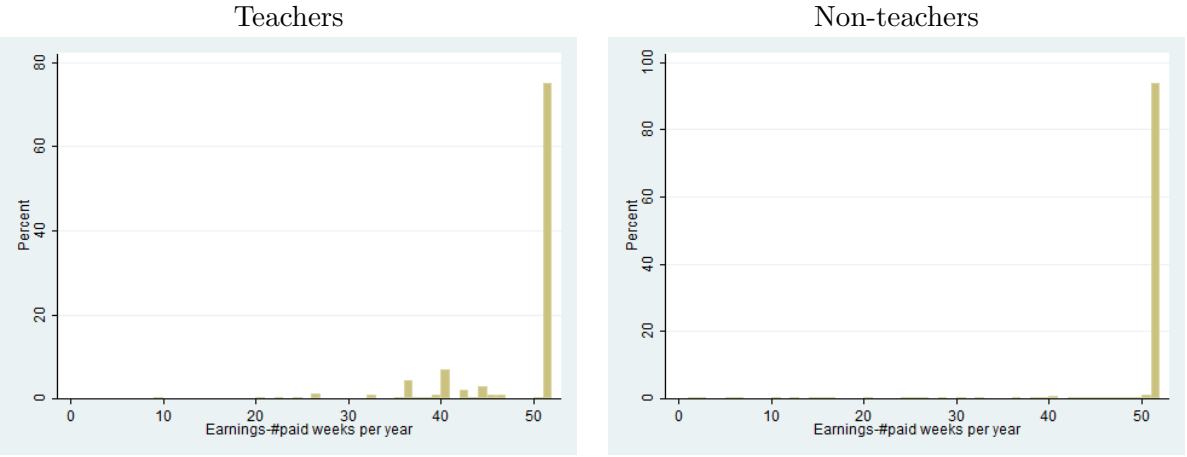
$$Wage_i = \frac{Earn_i}{Dhrs_i * 5} \tag{4.6}$$

where $Dhrs_i$ is the hours of work from each respondent's time diary. For this measure, I only include respondents sampled on the weekday and I multiply their daily hours by five. Thus, equation (4.6) has the advantage of using the individual's time diary hours

[25]The ATUS extract builder that I used to obtain ATUS data does not include the number of weeks worked variable so these figures are based on data from the 2010 CPS which I downloaded separately for this analysis. The sample includes all full time college educated workers in the January 2010 CPS.

[26]This may surprise some economists since spreading pay over the calendar year delays receipt of the final pay checks and thus decreases the net present value of the salary. Teachers may receive a benefit from spreading pay over the summer because they do not have to plan for uneven cash flows. Choosing to spread pay over the calendar year is rational if this benefit outweighs the small loss of potential interest income.

Figure 4.6: Weeks Worked

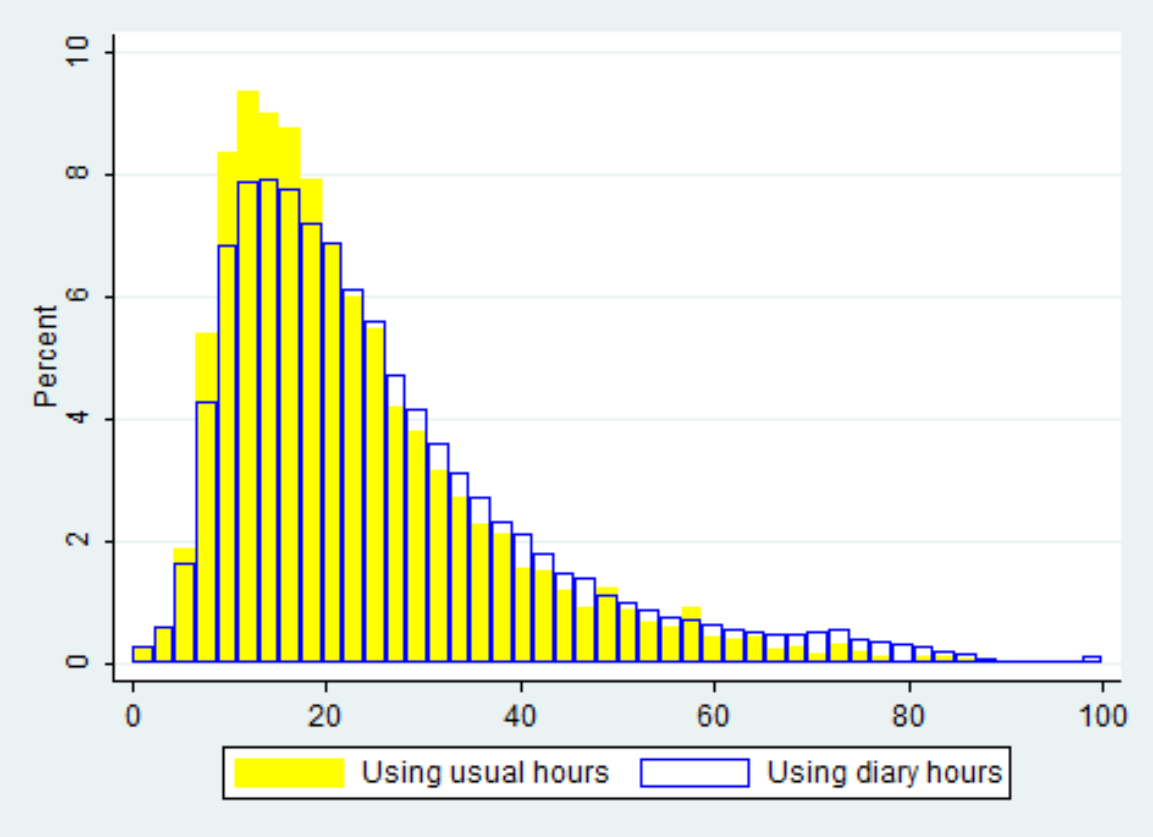


This figure shows the distribution of responses to the question about how many weeks the respondent worked. This question is only asked of respondents who report earnings annually. The ATUS extract builder that I used to obtain ATUS data does not include the number of weeks worked data so this figure is based on data from the 2010 CPS which I downloaded separately for this analysis. The sample includes all full time college educated workers in the January 2010 CPS. Teachers include elementary, middle, high school and special education teachers. Non-teachers include all other occupations.

rather than imposing the assumption that all respondents within an occupation work the same number of hours, but it replaces this with the assumption that respondents who work during the week work equal days and do not work at all on the weekend. Neither assumption is ideal, however, either one or the other is necessary since time diary data is only available for one day for each respondent. In the next section, results are shown for both and it is reassuring that they differ very little.

Figure 4.7 compares the distribution of hourly wages calculated using diary hours and usual hours of work, i.e. results of equation (4.4) and results from equation (4.5), for all full time college educated workers. The distributions are not that different. The mean hourly wage is slightly higher when calculated using diary hours of work. This reflects the fact that almost all workers over report their usual hours of work. In the next section I show that, while the overall distribution of hourly wages is not dramatically impacted by substituting $DHrs_o$ for $UHrs_i$, systematic biases by occupation make this adjustment very important when calculating the teacher wage gap. To preview this result, Figure 4.8 shows the same distribution for teachers. Notice that the shift in the distribution is much

Figure 4.7: Hourly Wages, Comparing Usual Hours of Work to Diary Hours of Work



This figure compares the distribution of hourly wages calculated using usual hours of work, i.e. equation (4.4), to hourly wages calculated using diary hours of work, i.e. equation (4.5), for all full time workers with at least a Bachelors degree and positive weekly earnings. I top code wages at $100/hour.

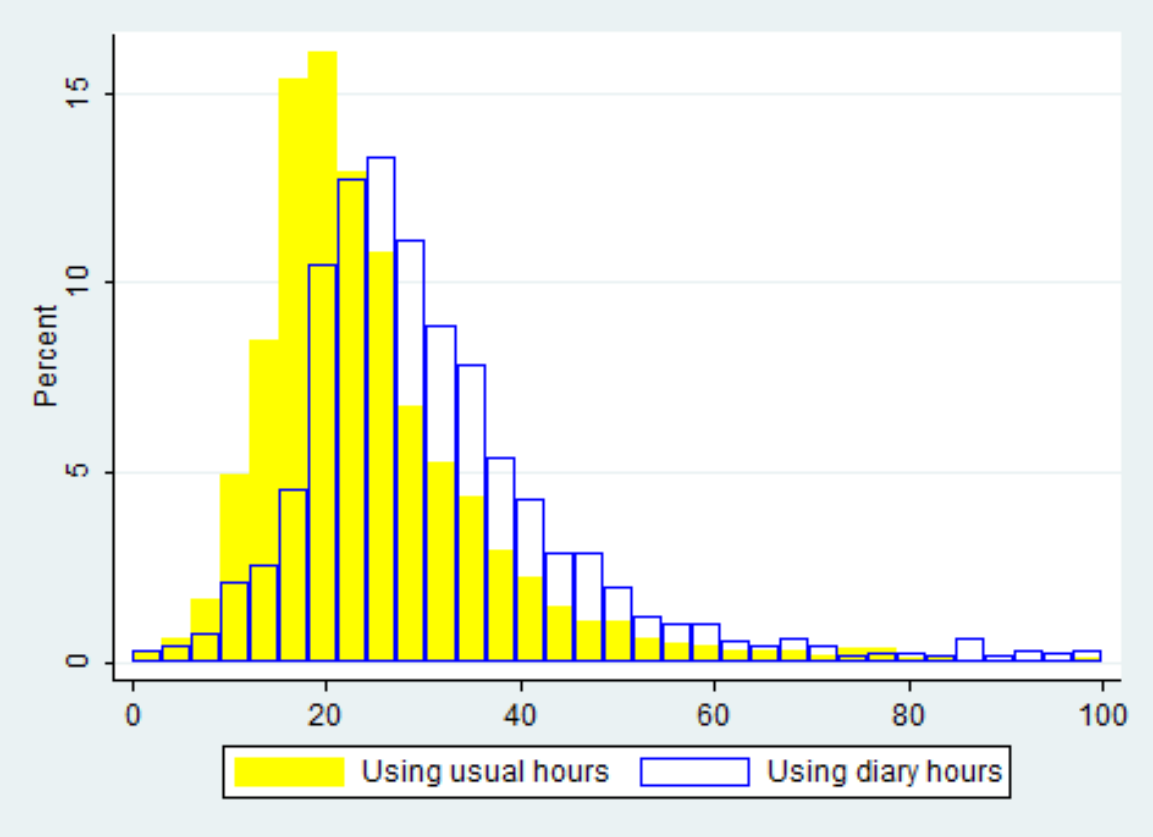
more obvious for teachers.

4.5.1 The Teacher Wage Gap Revisited

I estimate a simple Mincer style wage equation controlling for a vector of demographic variables, X_i, that includes a quadratic in age, additional education beyond a bachelors degree (a set of indicator variables for Masters, PhD and Professional degrees) and indicators for white and female.[27] To calculate the teacher wage gap, I include an indicator

[27] The omitted category is non-white males with no more than a Bachelors degree.

Figure 4.8: Hourly Wages of Teachers, Comparing Usual Hours of Work to Diary Hours of Work



This figure compares the distribution of hourly wages for teachers calculated using usual hours of work, i.e. equation (4.4), to hourly wages calculated using diary hours of work, i.e. equation (4.5), for all full time workers with at least a Bachelors degree and positive weekly earnings who are categorized as teachers at the time of their ATUS interview. I top code wages at $100/hour.

variable, $Teacher_i$, which equals one if the respondent is a teacher:

$$\ln(Wage_i) = \alpha + \beta' X_i + \gamma Teacher_i + \epsilon_i \tag{4.7}$$

The coefficient of interest, γ, is the difference in wages between teachers and other demographically similar workers, i.e. the teacher wage gap. I compare estimates of the teacher wage gap when $Wage_i$ is calculated using usual hours of work, equation (4.4), to estimates of the teacher wage gap when $Wage_i$ is calculated using diary hours of work, equation (4.5). Each observation is weighted using the weights provided by the ATUS and standard errors are calculated using the replicate weights provided by the ATUS.

Results are presented in columns (1) and (2) of Table 4.3. I find that when I use usual hours of work to calculate hourly wages, the teacher wage gap is 14.7%, i.e. teachers earn 14.7% less than demographically similar workers in other occupations. However, when I use diary hours of work to calculate hourly wages, I do not find a significant difference between teachers and non-teachers. That is, the teacher wage gap disappears.

I also report the teacher wage gap separately for elementary/middle school,secondary school and special education teachers, in columns (3) and (4) of Table 4.3. I do this is because secondary school teachers are less likely to over report their usual hours of work than are elementary/middle school or special education teachers. From a policy perspective, differentiating between secondary school, elementary/middle school and special education teachers is important because the license and training requirements differ for these groups, and the alternative labor market opportunities differ since most secondary school teachers have subject specific training that is more applicable to other sectors. I find that when I use diary hours of work to calculate hourly wages, there are significant differences between secondary school teachers and elementary/middle and special education teachers. Secondary school teachers earn an hourly wage that is 11% less than demographically similar workers. Elementary/middle school and special education teachers, on the other hand, do not earn hourly wages that are less than demographically similar workers.[28]

[28]The results suggest that special education teachers earn hourly wages that are almost 15% more than demographically similar workers but caution is advised when drawing conclusions based on this estimate. In particular, I believe it would be a bad policy to lower salaries for special education teachers based on this finding. Not only is the 95% confidence interval on the estimate quite wide (0.065, 0.229), the general

Table 4.3: The Teacher Wage Gap, Results for Hourly Wages Using Usual and Diary Hours of Work

	(1) Usual Hours	(2) Diary Hours	(3) Usual Hours	(4) Diary Hours
Teacher	-.147***	0.001		
	(0.016)	(0.016)		
Elementary/middle school			-.135***	0.033*
			(0.019)	(0.019)
Secondary school			-.209***	-.110***
			(0.025)	(0.024)
Special education			-.029	0.147***
			(0.039)	(0.042)
Age	0.062***	0.068***	0.062***	0.068***
	(0.004)	(0.004)	(0.004)	(0.004)
Age^2	-.0006***	-.0007***	-.0006***	-.0007***
	(0.00005)	(0.00005)	(0.00005)	(0.00005)
Masters degree	0.152***	0.171***	0.151***	0.171***
	(0.014)	(0.015)	(0.014)	(0.015)
PhD degree	0.217***	0.261***	0.216***	0.261***
	(0.025)	(0.025)	(0.025)	(0.025)
Professional degree	0.317***	0.355***	0.318***	0.357***
	(0.027)	(0.026)	(0.027)	(0.026)
White	0.02	0.045***	0.02	0.046***
	(0.013)	(0.014)	(0.013)	(0.014)
Female	-.160***	-.183***	-.162***	-.188***
	(0.011)	(0.011)	(0.011)	(0.012)
N	18157	18157	18157	18157
R^2	0.119	0.128	0.12	0.13

Dependent variable is ln(hourly wage) calculated from equation (4.4) using usual hours of work or equation (4.5) using diary hours of work. Significance *** 1%, ** 5%, * 10%. Each observation is weighted using the weights provided by the ATUS and standard errors are calculated with successive difference replication (SDR) variance estimation using the replicate weights provided by the ATUS. Sample includes ATUS respondents with at least a Bachelors degree who are full time workers with positive weekly earnings. Additionally 1,082 observations were dropped that did not have usual hours of work data or if there were too few respondents in their occupation to calculate a reliable measure for diary hours of work.

Table 4.4: The Teacher Wage Gap, Results for Hourly Wages Using Usual and Diary Hours of Work at the Individual Level

	(1) Usual Hours	(2) Diary Hours	(3) Usual Hours	(4) Diary Hours
Teacher	-.136***	-.028		
	(0.021)	(0.041)		
Elementary/middle school			-.128***	-.009
			(0.025)	(0.049)
Secondary school			-.194***	-.092
			(0.026)	(0.064)
Special Education			-.005	0.046
			(0.049)	(0.112)
Age	0.061***	0.07***	0.061***	0.07***
	(0.005)	(0.007)	(0.005)	(0.007)
Age^2	-.0006***	-.0007***	-.0006***	-.0007***
	(0.00006)	(0.00007)	(0.00006)	(0.00007)
008)				
Masters degree	0.155***	0.185***	0.154***	0.186***
	(0.017)	(0.024)	(0.017)	(0.024)
PhD	0.204***	0.327***	0.204***	0.327***
	(0.033)	(0.053)	(0.033)	(0.053)
Professional degree	0.329***	0.446***	0.33***	0.446***
	(0.035)	(0.05)	(0.035)	(0.05)
White	0.021	0.092***	0.021	0.093***
	(0.017)	(0.026)	(0.017)	(0.026)
Female	-.154***	-.153***	-.156***	-.156***
	(0.013)	(0.022)	(0.014)	(0.022)
Monday	0.03	0.027	0.03	0.027
	(0.023)	(0.036)	(0.023)	(0.036)
Tuesday	0.008	-.017	0.008	-.018
	(0.023)	(0.035)	(0.023)	(0.035)
Wednesday	-.002	-.063**	-.001	-.063**
	(0.022)	(0.03)	(0.022)	(0.03)
Thursday	-.012	-.045	-.012	-.045
	(0.029)	(0.034)	(0.029)	(0.034)
Holiday	-.018	1.679***	-.019	1.680***
	(0.058)	(0.272)	(0.058)	(0.271)
N	216549	215575	216549	215575
	8952	8017	8952	8017
R^2	0.12	0.096	0.121	0.096
	0.121	0.097	0.122	0.097

Dependent variable is ln(hourly wage) calculated from equation (4.6). Significance *** 1%, ** 5%, * 10%. Each observation is weighted using the weights provided by the ATUS and standard errors are calculated with successive difference replication (SDR) variance estimation using the replicate weights provided by the ATUS. Sample includes ATUS respondents with at least a Bachelors degree who are full time workers with positive weekly earnings who were interviewed on a weekday.

I also estimate a slight variation of equation (4.7) that uses each individual diary hours of work in place of averages by occupation. Recall that this has the advantage of allowing individuals to vary but limits the sample to weekday respondents only and imposes the assumption that they work five equal work days. In this case, I also include indicators that describe the reference day for the interview, Day_i, these include the day of the week and whether the reference day was a holiday.[29]

$$\ln(Wage_i) = \alpha + \beta^{'} X_i + \gamma Teacher_i + \delta Day_i + \epsilon_i \tag{4.8}$$

Table 4.4 reports these results which are very consistent with those in Table 4.3.[30] It is still the case that when using usual hours of work, teacher appear to be around 14% underpaid relative to similar workers in other occupations. When diary hours are used, however, teachers earn wages that are on par with similar workers in other occupations. The results are also similar when disaggregated by elementary/middle, secondary and special education. There is still evidence that secondary school teachers are underpaid on the order of magnitude of 10%, however, this result is no longer statistically significant at conventional levels. As before, elementary/middle school and special education teachers do not earn hourly wages that are less than demographically similar workers.[31]

Tables 7.11 - 7.15 in the appendix examine this result in more detail. First, Tables 7.11 and 7.12 report results for equation (4.7) for women and men separately. This is important since wage discrimination against teachers (especially elementary/middle school teachers) may be obscured by occupational segregation. I find that, when using diary hours of work, female elementary/middle school teachers are paid hourly wages that are 6% *more* than

consensus is that most districts face a shortage of special education teachers so lower salaries would be counter productive. It is likely that special education teaching is a different job than elementary/middle or secondary school teaching on a number of unmeasured dimensions. The difference in pay may represent a compensating differential for a less desirable assignment.

[29] The omitted category is a non-holiday Friday.

[30] The negative coefficients on the days of week variables are as expected. The omitted category is Friday. The negative coefficients on all other days of the week show that, on average, people work more hours on these days than they do on Fridays, and thus earn a lower hourly wage when a non-Friday is used to calculate their hourly wage. Likewise, the positive coefficient on *Holiday* is as expected. The fact that the descriptors of the reference day are insignificant when hourly wages are calculated using usual hours of work shows that responses to this question do not vary by day of the interview. This is reassuring.

[31] The surprising result for special education teachers discussed in the previous footnote, is no longer evident. This lends further support to my assertion that it would be a mistake to infer that special education teachers are overpaid.

females in other occupations and female secondary teachers are paid hourly wages similar to other females in other occupations. Male elementary/middle school and secondary school teachers are paid hourly wages that are 7% and 20%, respectively, *less* than males in other occupations. Tables 7.13 and 7.14 report results separately for respondents whose highest degree is a Bachelors and respondents whose highest degree is a Masters. For both levels of education, the results are consistent with Table 4.3. Table 7.15 differentiates between public school teachers and private school teachers. I find that the average private school teacher is paid less than demographically similar workers, even after accounting for diary hours of work. This result holds for both elementary/middle and secondary private school teachers.

Each of the aforementioned subgroup results assumes that teachers all work the same number of hours, that is $Dhrs_o$ is constant across gender, education and sector.[32] It is possible to calculate $Dhrs_o$ separately by subgroup, however, the sample size on which the estimates are based falls and more measurement error is introduced. Noting this caveat, I re-estimate equation 4.7 using subgroup specific $Dhrs_o$ to calculate $Wage_i$.

In Table 4.5 reports the results. The first column reproduces column (2) from Table 4.3 for comparison. The second and third columns limit the sample to female and male respondents respectively. I find that, on average, while female teachers are paid hourly wages that are comparable to demographically similar females in other occupations, male teachers are paid 13% less than demographically similar males in other occupations. The fourth and fifth columns limit the sample to respondents whose highest degree is a Bachelors and Masters respectively. I find that, on average, teachers with a Masters are paid 6% less than demographically similar workers with Masters degrees in other occupations. Finally, the sixth column disaggregates teachers into public and private sector workers. Here I find that, in contrast to public school teachers, private school teachers are paid hourly wages that are 20% less than demographically similar workers in other occupations.

Similarly, Table 4.6 reports the results using individual diary hours times five for weekday respondents. The first column reproduces column (2) from Table 4.4 for comparison. In this specification, I find that male teachers are paid 18% less than demographically similar males in other occupations. The result for teachers with Masters degrees still indicates

[32]Therefore, the subgroup results are driven by differences in $Earn_i$.

Table 4.5: The Teacher Wage Gap, Results for Hourly Wages Using Diary Hours of Work, by Subgroup

	(1) Full Sample	(2) Female	(3) Male	(4) Bachelors	(5) Masters	(6) By Sector
Teacher	0.0009	0.026	-.128***	0.027	-.061**	
	(0.017)	(0.018)	(0.03)	(0.02)	(0.025)	
Public teacher						0.032*
						(0.017)
Private teacher						-.201***
						(0.039)
Age	0.069***	0.061***	0.073***	0.072***	0.049***	0.068***
	(0.004)	(0.006)	(0.005)	(0.005)	(0.008)	(0.004)
Age^2	-.0007***	-.0006***	-.0007***	-.0007***	-.0005***	-.0007***
	(0.00005)	(0.00007)	(0.00006)	(0.00006)	(0.00009)	(0.00005)
Masters degree	0.17***	0.231***	0.123***			0.171***
	(0.015)	(0.018)	(0.023)			(0.015)
PhD degree	0.259***	0.312***	0.221***			0.261***
	(0.025)	(0.052)	(0.027)			(0.025)
Professional degree	0.354***	0.395***	0.318***			0.355***
	(0.026)	(0.051)	(0.027)			(0.026)
White	0.044***	0.012	0.071***	0.052***	0.019	0.046***
	(0.014)	(0.018)	(0.023)	(0.017)	(0.025)	(0.014)
Female	-.185***			-.212***	-.134***	-.183***
	(0.012)			(0.015)	(0.023)	(0.011)
N	18157	8945	9212	11394	4954	18157
R^2	0.127	0.105	0.094	0.109	0.041	0.13

Dependent variable is ln(hourly wage) calculated from equation (4.5) using diary hours of work. Significance *** 1%, ** 5%, * 10%. Each observation is weighted using the weights provided by the ATUS and standard errors are calculated with successive difference replication (SDR) variance estimation using the replicate weights provided by the ATUS. Column (1) replicates the result from Table 4.3 for comparison. Columns (2)-(5) limit the sample to only female, male, highest degree is a BA, highest degree is an MA respectively. Column (6) repeats column (1) disaggregating teachers into public and private sector workers.

Table 4.6: The Teacher Wage Gap, Results for Hourly Wages Using Diary Hours of Work at the Individual Level, by Subgroup

	(1) Full Sample	(2) Female	(3) Male	(4) Bachelors	(5) Masters	(6) By Sector
Teacher	-.028	0.036	-.180***	-.005	-.060	
	(0.041)	(0.048)	(0.059)	(0.053)	(0.06)	
Public teacher						-.013
						(0.043)
Private teacher						-.124
						(0.089)
Age	0.07***	0.065***	0.075***	0.078***	0.041***	0.07***
	(0.007)	(0.01)	(0.007)	(0.007)	(0.016)	(0.007)
Age^2	-.0007***	-.0007***	-.0007***	-.0008***	-.0004**	-.0007***
	(0.00007)	(0.0001)	(0.00008)	(0.00008)	(0.0002)	(0.00007)
Masters degree	0.185***	0.237***	0.137***			0.185***
	(0.024)	(0.039)	(0.03)			(0.024)
PhD degree	0.327***	0.325***	0.318***			0.327***
	(0.053)	(0.097)	(0.065)			(0.053)
Professional degree	0.446***	0.51***	0.399***			0.446***
	(0.05)	(0.071)	(0.072)			(0.05)
White	0.092***	0.07*	0.112***	0.089***	0.1**	0.093***
	(0.026)	(0.038)	(0.034)	(0.03)	(0.046)	(0.026)
Female	-.153***			-.187***	-.075**	-.153***
	(0.022)			(0.029)	(0.038)	(0.022)
Monday	0.027	-.007	0.054	0.043	0.021	0.027
	(0.036)	(0.048)	(0.054)	(0.045)	(0.061)	(0.036)
Tuesday	-.017	0.045	-.073	0.006	-.032	-.018
	(0.035)	(0.047)	(0.046)	(0.044)	(0.059)	(0.035)
Wednesday	-.063**	-.075*	-.053	-.060*	-.061	-.065**
	(0.03)	(0.043)	(0.042)	(0.036)	(0.06)	(0.03)
Thursday	-.045	-.038	-.050	-.011	-.106**	-.046
	(0.034)	(0.046)	(0.043)	(0.042)	(0.054)	(0.035)
Holiday	1.679***	1.945***	1.397***	1.583***	1.797***	1.677***
	(0.272)	(0.432)	(0.337)	(0.337)	(0.479)	(0.272)
N	215575	216008	216116	215888	216298	215575
R^2	0.096	0.09	0.086	0.079	0.054	0.096

Dependent variable is ln(hourly wage) calculated from equation (4.6). Significance *** 1%, ** 5%, * 10%. The sample is limited to respondents whose ATUS interview was on a weekday. Each observation is weighted using the weights provided by the ATUS and standard errors are calculated with successive difference replication (SDR) variance estimation using the replicate weights provided by the ATUS. Column (1) replicates the result from Table 4.3 for respondents with an ATUS interview on a weekday. Columns (2)-(5) limit the sample to only female, male, highest degree is a BA, highest degree is an MA respectively. Column (6) repeats column (1) disaggregating teachers into public and private sector workers.

that they earn on the order of 6% less than demographically similar workers with Masters degrees in other occupations, however the result is no longer statistically significant at conventional levels. Likewise, the result for private school teachers still shows that they are paid hourly wages that are less than demographically similar workers in other occupations, but this result is no longer statistically significant at conventional levels.

These subgroup results provide a slightly more nuanced story to complement the main result of this paper. The fact that a wage gap that exists for male teachers is likely because male non-teachers are concentrated in higher paying occupations than female non-teachers. It suggests that lower wages for teachers are, at least in part, due to occupational segregation. The fact that a wage gap may exist for teachers with Masters degrees and for private school teachers provides evidence that Masters degrees in education are not rewarded at the same rate as Masters degrees in other occupations and that private school wages are below the market wage.[33]

4.5.2 Other Wage Gaps Revisited

Lastly, I compare estimates of wage gaps by education, gender and race without an indicator for $Teacher_i$. Here I add workers with at least a high school diploma but less than a Bachelors degree back into the sample and collapse the indicators for levels of education higher than a Bachelors degree into one variable, $Masters_i$, which should now be interpreted as having at least a Masters degree.[34] Specifically, I estimate:

$$\ln(Wage_i) = \alpha + \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 Bachelors_i + \beta_4 Masters_i + \beta_5 White_i + \beta_6 Female_i + \epsilon_i \tag{4.9}$$

I focus on β_3, the college-high school wage gap, β_5 the white-non white wage gap, and β_6 the female-male wage gap.[35] As before, each observation is weighted using the

[33] Private school wages that are below the market wage are easily explained by compensating differentials.

[34] The omitted category is now non-white males with at least a high school diploma but less than a Bachelors degree.

[35] Specifically, these are respectively: the wage premium for full-time workers who complete a Bachelors degree but no more over full-time workers with only a high school diploma or a high school diploma and some college; the wage premium(penalty) for full-time female workers with at least a high school diploma over full-time male workers with at least a high school diploma and; the wage premium for full-time white workers with at least a high school diploma over full-time non-white workers with at least a high school diploma.

Table 4.7: Other Wage Gaps, Results for Hourly Wages Using Usual and Diary Hours of Work

	(1) Usual Hours	(2) Diary Hours
Age	0.061***	0.067***
	(0.002)	(0.002)
Age^2	-.0006***	-.0007***
	(0.00002)	(0.00002)
Bachelors degree	0.449***	0.476***
	(0.008)	(0.008)
Masters degree (or higher)	0.628***	0.691***
	(0.011)	(0.011)
White	0.067***	0.08***
	(0.009)	(0.009)
Female	-.185***	-.195***
	(0.006)	(0.006)
N	48561	48561
R^2	0.265	0.284

Dependent variable is ln(hourly wage) calculated from equation (4.4) using usual hours of work or equation (4.5) using diary hours of work. Significance *** 1%, ** 5%, * 10%. Each observation is weighted using the weights provided by the ATUS and standard errors are calculated with successive difference replication (SDR) variance estimation using the replicate weights provided by the ATUS. Sample includes ATUS respondents with at least a high school diploma who are full time workers with positive weekly earnings. Observations were dropped that did not have usual hours of work data or if there were too few respondents in an occupation to calculate a reliable measure for diary hours.

weights provided by the ATUS and standard errors are calculated with successive difference replication (SDR) variance estimation using the replicate weights provided by the ATUS. I assume that diary hours of work are consistent within occupation and do not vary within occupation by level of education, race or gender.

Results are presented in Table 4.7. I find that the college-high school, the female-male and the white-non white wage gaps are all slightly wider when diary hours of work are used to calculate hourly wages than when usual hours of work are used to calculate hourly wages.[36] The college-high school wage gap increases from 0.449 to 0.476 (a 6% increase).

[36]Table 4.3 also shows widening education, gender and racial wage gaps amongst workers with at least a Bachelors degree.

Table 4.8: Other Wage Gaps, Results for Hourly Wages Using Usual and Diary Hours of Work at the Individual Level

	(1) Usual Hours	(2) Diary Hours
Age	0.06***	0.062***
	(0.002)	(0.003)
Age^2	-.0006***	-.0006***
	(0.00003)	(0.00004)
Bachelors degree	0.451***	0.507***
	(0.011)	(0.015)
Masters degree (or higher)	0.624***	0.744***
	(0.014)	(0.02)
White	0.063***	0.107***
	(0.012)	(0.016)
Female	-.184***	-.176***
	(0.008)	(0.013)
Monday	0.007	-.008
	(0.013)	(0.02)
Tuesday	0.002	-.034
	(0.013)	(0.021)
Wednesday	0.003	-.049***
	(0.011)	(0.017)
Thursday	-.007	-.042**
	(0.014)	(0.02)
Holiday	-.028	0.95***
	(0.033)	(0.21)
N	216549	213361
R^2	0.264	0.187

Dependent variable is ln(hourly wage) calculated from equation (4.4) using usual hours of work or equation (4.5) using diary hours of work with $DHrs_i$ in place of $DHrs_o$. Significance *** 1%, ** 5%, * 10%. Each observation is weighted using the weights provided by the ATUS and standard errors are calculated with successive difference replication (SDR) variance estimation using the replicate weights provided by the ATUS. Sample includes ATUS respondents with at least a high school diploma who are full time workers with positive weekly earnings who were interviewed on a weekday.

The female-male wage gap increases from -0.185 to -0.195 (also a 6% increase).[37] And the white-non white wage gap increases from 0.067 to 0.080 (a 16% increase).[38]

As before, I also present results using individual diary hours and a vector of indicators that describe the reference day for the interview. These results are in Table 4.8 and echo the findings reported in table 4.7 with the exception of the finding regarding the female-male wage gap. In this specification, the use of diary hours *decreases* the female-male wage gap from -0.184 to -0.176.

4.6 Discussion

Teachers, like all workers, make labor market decisions based on a complex set of information. They consider each job as a bundle of characteristics including amenities such as location, work schedule and their level of personal satisfaction. Non-pecuniary factors certainly play a role in a college graduate's decision to pursue and subsequently remain in teaching. However, to ignore the obvious impact that compensation has on the number and quality of potential teachers would be a mistake. If teaching pays less than other occupations, schools will be left with a small pool of lower quality candidates to choose from. If teaching pays more than other occupations, there will be a surplus of potential teachers.

To properly compare wages for teachers to wages in other occupations, researchers must take care to properly account for hours of work. The time diary data in the ATUS provide a unique opportunity to move beyond warring anecdotes about teachers who work short contract days/years and teacher who work extraordinary hours staying late to provide extra help, taking work home to grade and planning and attending extra curricular activities on the weekend. Time diaries provide a clearer picture of hours of work for teachers and

[37] These magnitudes are consistent with Frazis and Stewart (2004) who find that accounting for over reporting increases the college-high school wage gap by 4.1% and the female-male wage gap by 5.4%.

[38] Wald tests show that these differences are all statistically significant at the 1% level. To conduct these tests I stack columns (1) and (2) into one regression. Specifically, I duplicate the data frame so that each observation is repeated and estimate $y = D + (1-D) + \beta_1 Age_i * D + \alpha_1 Age_i * (1-D) + ... + \beta_6 Female_i * D + \alpha_6 Female_i * (1-D) + \epsilon_i$ where $D = 0$ for the initial observation and $D = 1$ for the duplicate observation and $y = Wage_i$ using usual hours of work for the initial observation and $y = Wage_i$ using diary hours of work for the duplicate observation. I then test the hypotheses that $\beta_3 = \alpha_3$ for the college-high school wage gap, $\beta_5 = \alpha_5$ for the white-non white wage gap and $\beta_6 = \alpha_6$ for the female-male wage gap. Each of these null hypotheses are rejected at the 1% level.

non-teachers than either administrative data or recall data from surveys that ask about a usual or typical week of work. Using the ATUS, I find that teachers work more than they are contractually required to. They bring work home, work on the weekends during the summer months. However, teachers work less than they self-report when asked to recollect a usual week of work. All workers work less than they self-report when asked to recollect a usual week but teachers are more likely to over report their hours than are workers in other occupations.

I construct measures of diary hours per week by occupation and find that teachers work an average of 38.0 hours/week during the school year and 21.5 hours/week during the summer. When I use this measure rather than a measure of usual hours of work per week, I find that teachers' hourly wages are no more and no less than workers in other occupations. In many ways this is exactly as economists would predict and is confirmation of the power of market forces to set wages across occupations.

Additionally, I find that averaging across all teachers obscures some interesting detail. Secondary school teachers appear to have notably different work schedules and are less likely to over report their usual hours of work than elementary/middle school and special education teachers. When diary measures of hours of work are used to calculate hourly wages, secondary school teachers earn 11% less than demographically similar workers in other occupations while elementary, middle and special education teachers are not underpaid relative to other occupations.[39] I take this as evidence that policy makers should consider abandoning the single salary schedule that forces districts to pay all teacher the same regardless of the grade and subject they teach. Licensing and alternative labor market opportunities differ for these categories of teachers and I have shown that work schedules differ as well. With the current system, it is likely that schools face a surplus of elementary school teachers and a shortage of high quality secondary school teachers. Raising secondary school teacher wages could help alleviate this problem.[40]

[39] Male elementary school teachers earn hourly wages that are approximately 7% less than other males with at least a Bachelors degree. If occupational segregation accounts for the teacher wage gap, then this indicates that elementary teachers are also underpaid relative to other occupations.

[40] Research indicates that it is in fact misleading to talk about a single teacher labor market. Many districts simultaneously face a shortage of math, science and special education teachers at all levels and a surplus of elementary, English and social studies teachers. This suggests that further disaggregating by subject may be important as well. Unfortunately, ATUS data only allows for disaggregation by elementary/middle, secondary and special education.

Future research should attempt a more detailed disaggregation of teachers to look for regional or demographic difference in time use patterns and hourly wages. Ideally, a more detailed disaggregation would also include separating elementary and middle school teachers and differentiating between subjects at the secondary school level but this is not possible with the data in the ATUS. Other potential avenues for future research include using the timing and location of work information available in the ATUS to explore the fact that teachers may be better able to align their work day and year with their own children's school day and year. This non-pecuniary benefit may be of interest especially if it enables teachers to spend less on paid childcare than other workers. Policy debates over reforms aimed at increasing the quality of the teacher labor force will only benefit from more accurate and more detailed information about teacher time use and wages like that provided in this study.

Chapter 5

Conclusion

The three essays that comprise this dissertation each address different aspects of teacher labor markets. The reader, however, will undoubtedly notice many common themes. Most obvious is that all of the essays investigate issues related to teacher contracts. Teacher contracts are one of, if not *the* most influential and policy mailable tools we have to influence teacher productivity. It is crucial that theory and evidence, such as that provided herein, developed using the tools of labor economics be an integral part of the policy discussions. Reforms rooted in economic theory and evidence have a much higher likelihood of success than do ad-hoc attempts based on gut feelings or politics.

The current push for reform, both locally and nationally, has largely focused on pay for performance. In the first part of this dissertation, I present a theoretical model that provides clarity regarding the trade-offs inherent in moving from a contract that bases compensation on inputs to one that bases compensation on outputs. This model shows that in order for output based pay for performance to be effective, districts must also provide support for decentralized decision making. I use data from a local program, Minnesota's Q-Comp, to test this but am unable to offer conclusive evidence.

In the second part of this dissertation, I also distinguish between compensation based on inputs and compensation based on outputs. I provide important new survey data about the prevalence of input and output based pay for performance and show that unions are not the barrier to reforms that many assert they are. Although unions are historically associated with the single salary schedule, there is evidence that both input and output

based compensation reforms are politically feasible in both unionized and non-unionized districts.

Taking these two essays together, I suspect that unions likely will be more supportive of pay for performance if it is paired with management support for decentralized decision making, i.e. additional time for teacher collaboration. While empirical backing for this assertion is left for future research, I believe that pay for performance will garner more union support when thoughtfully crafted in a way that respects teachers' professionalism and local knowledge rather than imposed as purely an accountability system with the implied accusation that teachers are not working hard enough.

In the third part of this dissertation, I directly address the question of how hard teachers work. I find that teachers work an average of 34.5 hours per week. This includes 38.5 hours per week during the school year and 21.5 hours per week during the summer months. It is clearly problematic to assume that teachers do not work beyond what is contractually required. It is also problematic, however, to rely on teacher self reports regarding their usual hours of work since teachers are more prone to over reporting than workers in other occupations.

This third essay is not as clearly linked to the first two, however, pay for performance will likely change the number of hours that teachers work and their final salary – each of these will impact the hourly wage that teachers earn. If there were a way to link teacher time diary data to student outcomes, it would be possible to investigate whether more productive teachers are currently working more or fewer hours than their less productive peers. If they are currently working fewer hours, they are already earning a higher effective wage. If, on the other hand, they are currently working longer hours, then pay for performance may equalize wages across teachers of varying degrees of productivity.

The third essay also relates to the first two essays because calls for differentiated pay based on factors other than experience and education. I find that secondary school teachers earn hourly wages that are less than demographically similar workers in other occupations, but this is not the case for elementary/middle school teachers. Differentiating pay based on assignment is not the same as differentiating pay based on output but it is certainly a reform in the same vein.

5.1 Plans for Future Research

As is so often the case, the process of answering the research questions posed in this dissertation lead to still more questions. Below is a brief summary possible research strands to follow up on some of the unanswered questions alluded to throughout.

First, to follow up on the complementarity of pay for performance and delegated decision making, it would be interesting to test the theory with new and different data. The analysis presented herein does not refute the theory offered, but it falls short of supporting the theory. Two possible sources for new data include updated and more detailed data from Minnesota's Q-Comp and data from a similar program in Texas. If either, or both, of these provide clearer measures of support for decentralized decision making, I would be better able to test of the theory's predictions.

Second, to follow up on the role that unions play in education reform, particularly compensation and tenure reform, it would be interesting to make better use of the unbalanced panel nature of the SASS rather than simply treating it as a pooled cross-section. Specifically, there are over 2,000 districts that are sampled in both 2003-04 and 2007-08 and future waves of the SASS will extend this panel. Some districts changed in the union status between 2003-04 and 2007-08 and the recent political turmoil in states such as Wisconsin should only serve to generate more changes. It would be interesting to use the SASS, and possibly future iterations of my survey, to investigate the impact that changes unionization status and change in state laws regarding collective bargaining have on teacher contracts and teacher labor markets.

Third, the ATUS provides an opportunity to look at teacher time use patterns across the day and year. It may be that teachers time use patterns, more so than their total number of hours, is a compensating differential. Specifically, teachers with young children may like the fact that they are able to work hours that coincide with their children's school day. It is also possible that a teacher's schedule is a negative for others, particularly those without children who would like to work different hours. Disaggregating the data based on respondents' number and age of children and also on gender, marital status and spouses' earnings and time use may prove interesting.

Additionally, the areas of overlap between the three essays raise a number of questions such as: (1) Are unions more or less supportive of pay for performance when it is coupled

with management reforms? (2) Are highly productive teachers (i.e. those who would earn output based pay for performance bonuses) allocating their time differently than less productive teachers?[1] (3) Do pay for performance and management reforms change how teachers allocate their time? (4) Do teachers in unionized districts allocate their time differently than those in non-unionized districts? Clearly, this dissertation is only the start of a long research agenda to try and understand teacher labor markets and educational production more broadly.

[1] Although it is not possible to answer this with the ATUS, other time diary data might be available.

Chapter 6

Bibliography

(2012). *American Time Use Survey User's Guide: Understanding ATUS 2003 to 2011*. http://www.bls.gov/tus/atususersguide.pdf.

Aaronson, D., Barrow, L., and Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics*, 25(1):95–135.

Abraham, K., Flood, S., Sobek, M., and Thorn, B. (2011). American Time Use Survey Data Extract System: Version 2.4 [Machine-readable database]. http://www.atusdata.org.

Abraham, K., Maitland, A., and Bianchi., S. (2006). Nonresponse in the American Time Use Survey: Who is missing from the data and how much does it matter? *Public Opinion Quarterly*, 70:676–703.

Aghion, P. and Tirole, J. (1997). Formal and real authority in organizations. *Journal of Political Economy*, 105(1):1–29.

Allegretto, S., Corcoran, S., and Mishel, L. (2004). How does teacher pay compare? Methodological challenges and answers. Economic Policy Institute.

Allegretto, S., Corcoran, S., and Mishel, L. (2008). The teaching penalty: Teacher pay losing ground. Economic Policy Institute.

Allen, S. (1988). *Unions and Job Security in the Public Sector*, chapter in When Public

Sector Workers Unionize. Eds. R. Freeman and C. Ichniowski. University of Chicago Press. available at: http://www.nber.org/books/free88-1.

Atkinson, A., Burgess, S., Croxson, B., Gregg, P., Propper, C., Slater, H., and Wilson, D. (2004). Evaluating the impact of performance-related pay for teachers in England. *Department of Economics, University of Bristol, UK, The Centre for Market and Public Organisation*, 60.

Bacolod, M. (2007). Do alternative opportunities matter? the role of female labor markets in the decline of teacher quality. *Review of Economics and Statistics*, 89:737–51.

Baker, G., Gibbons, R., and Murphy, K. (1994). Subjective performance measures in optimal incentive contracts. *The Quarterly Journal of Economics*, 109(4):1125–1156.

Ballou, D. (2001). Pay for performance in public and private schools. *Economics of Education Review*, 20(1):51–61.

Ballou, D. and Podgursky, M. (2002). Returns to seniority among public school teachers. *The Journal of Human Resources*, 37(4):892–912.

Baum-Snow, N. and Neal, D. (2009). Mismeaurement of usual hours worked in the Census and ACS. *Economic Letters*, 102(1):39–41.

Belfield, C. and Heywood, J. (2008). Performance pay for teachers: Determinants and consequences. *Economics of Education Review*, 27(3):243–252.

Blanchflower, D. and Bryson, A. (2007). *What effect do unions have on wages now and would Freeman and Medoff be surprised?*, chapter in What Do Unions Do? A twenty year perspective. Eds. J. Bennet and B. Kaufman. Transaction Publishers. Originally published in a six-part symposium in the Journal of Labor Research.

Bloom, N. and Van Reenen, J. (2011). Human resource managament. *Handbook of Labor Economics*, 4b:1697–1763.

Budd, J. (2007). *The effect of unions on employee benefits and non-wage compensation: monopoly power, collective voice, and facilitation*, chapter in What Do Unions Do? A

twenty year perspective. Eds. J. Bennet and B. Kaufman. Transaction Publishers. Originally published in a six-part symposium in the Journal of Labor Research.

Corcoran, S., Evans, W., and Schwab, R. (2004). Women, the labor market,and the declining relative quality of teachers. *Journal of Policy Analysis and Management*, 23:449–70.

Drago, R., Caplan, R., Costanza, D., Brubaker, T., Cloud, D., Harris, N., Kashian, R., and Riggs, T. L. (1999). New estimates of working time for elementary school teachers. *Monthly Labor Review*, (4):31–40.

Eberts, R. (2007). Teachers unions and student performance: Help or hindrance? *The Future of Children*, 17(1):175–200.

Figlio, D. and Winicki, J. (2005). Food for thought: the effects of school accountability plans on school nutrition. *Journal of Public Economics*, 89(2-3):381–394.

Flyer, F. and Rosen, S. (1997). The new economics of teachers and education. *Journal of Labor Economics*, 15(1):S104–S139.

Frazis, H. and Stewart, J. (2004). What can time-use data tell us about hours of work? *Monthly Labor Review*, 127(12):3–9.

Freeman, R. (1986). Unionism comes to the public sector. *Journal of Economic Literature*, 24:41–86.

Freeman, R. and Medoff, J. (1984). *What Do Unions Do?* Basic Books Inc.

Fryer, R. (2011). Teacher incentives and student achievement: Evidence from New York City Public Schools. Technical report, National Bureau of Economic Research Working Paper No. 16850.

Goldhaber, D., DeArmond, M., Player, D., and Choi, H. (2008). Why do so few public school districts use merit pay? *Journal of Education Finance*.

Goldhaber, D. and Liu, A. Y.-H. (2002). Occupational choices and the academic proficiency of the teacher workforce. developments in school finance 2001-02. National Center for Education Statistics, U.S. Department of Education.

Grossman, S. and Hart, O. (1986). The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy*, 94(4):691–719.

Gunderson, M. (2007). *Two face of union voice in the public sector*, chapter in What Do Unions Do? A twenty year perspective. Eds. J. Bennet and B. Kaufman. Transaction Publishers. Originally published in a six-part symposium in the Journal of Labor Research.

Hanushek, E. (2003). The failure of input-based resource policies. *The Economic Journal*, 113(485):F64–F98.

Hanushek, E. and Pace, R. (1995). Who chooses to teach (and why)? *Economics of Education Review*, 14(2):101–117.

Hanushek, E. and Rivkin, S. (2006). *Handbook of the Economics of Education*, volume 2, chapter Teacher Quality, page Chapter 18. Elsevier.

Harris, D. and Adams, S. (2007). Understanding the level and causes of teacher turnover: A comparison with other professions. *Economics of Education Review*, 26(3):325–337.

Holmes, A. (1979). Union activity and teacher salary structure. *Industrial Relations*.

Holmstrom, B. and Milgrom, P. (1991). Multitask principal–agent analyses: incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and organization*, 7(special issue):24.

Hoxby, C. (1996). How teachers' unions affect education production. *The Quarterly Journal of Economics*.

Hoxby, C. M. and Leigh, A. (2004). Pulled away or pushed out? explaining the decline of teacher aptitude in the united states. *American Economic Review*, 94:236–40.

Ittner, C. and Larcker, D. (2002). Determinants of performance measure choices in worker incentive plans. *Journal of Labor Economics*.

Jacob, B. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the chicago public schools. *Journal of Public Economics*, 89(5-6):761–796.

Jacob, B. and Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, 26(1):101.

Jacob, B. and Levitt, S. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, 118(3):843.

Juster, F. T. and Stafford, F. (1991). The allocation of time: Empirical findings, behavioral models, and problems of measurement. *Journal of Economic Literature*, 29:471–522.

Kaufman, B. (2007). *What unions do: Insights from economic theory*, chapter in What Do Unions Do? A twenty year perspective. Eds. J. Bennet and B. Kaufman. Transaction Publishers. Originally published in a six-part symposium in the Journal of Labor Research.

Keigher, A. (2010). Teacher attrition and mobility: Results from the 2008-09 Teacher Follow-up Survey (NCES 2010-353). Technical report, U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Krantz-Kent, R. (2008). Teachers' work patterns: When, where, and how much do U.S. teachers work? *Monthly Labor Review*, (3):52–9.

Lincove, J. A. (2012). Can teacher incentive pay improve student performance on standardized tests. Unpublished manuscript. Presented at the American Education Finance and Policy conference.

Lockwood, J. R., McCaffrey, D., Hamilton, L., Stecher, B., Le, V., and Martinez, J. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*.

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

Lovenheim, M. (2009a). The effect of teachers' unions on education production: Evidence from union election certifications in three midwestern states. *Journal of Labor Economics*.

Lovenheim, M. (2009b). The effect of teachers' unions on education production: Evidence from union election certifications in three midwestern states. *Journal of Labor Economics*.

Martins, P. (2009). *Individual teacher incentives, student achievement and grade inflation*. IZA Discussion Paper No. 4051.

Murnane, R. and Cohen, D. (1986). Merit pay and the evaluation problem: Why most merit pay plans fail and few survive. *Harvard Educational Review*, pages 1–18.

Neal, D. (2011). The design of performance pay in education. *Handbook of the Economics of Education*, 4:495–550.

Nelson, H. and Podgursky, M. (2003). Correspondence. *Education Next*, 3(4):5–6.

Nye, B., Konstantopoulos, S., and Hedges, L. (2004). How large are teacher effects. *Educational Evaluation and Policy Analysis*, 26(3):237–257.

Podgurksy, M. (2003). Fringe benefits. *Education Next*, 3(3):71–6.

Podgursky, M. and Mishel, L. (2005). National Council on Teacher Quality http://www.nctq.org/p/publications/docs/nctq$_s quare_o f f_2$0071202080402.*pdf*.

Podgursky, M., Monroe, R., and Watson, D. (2004). The academic quality of public school teachers: An analysis of entry and exit behavior. *Economics of Education Review*, 23(5):507–518.

Podgursky, M. and Tongrut, R. (2006). (Mis-)measuring the relative pay of public school teachers. *Education Finance and Policy*, 1(4):425–40.

Prendergast, C. (2002). The tenuous trade-off between risk and incentives. *Journal of Political Economy*, 110(5).

Radner, R. (1993). The organization of decentrialzed information processing. *Econometrica*, 61(5):1109–1146.

Richwine, J. and Biggs, A. (2011). Assessing the compensation of public-school teachers. Heritage Center.

Rivkin, S., Hanushek, E., and Kain, J. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458.

Robinson, J. and Bostrom, A. (1994). The overestimated workweek? What time diary measures suggest. *Monthly Labor Review*, (8):11–23.

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

Rockoff, J., Jacob, B., Kane, T., and Staiger, D. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy*, 6(1):43–74.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, 25(1).

Ruggles, S., Alexander, J. T., Genadek, K., Goeken, R., Schroeder, M. B., and Sobek., M. (2010). Integrated public use microdata series: Version 5.0 [machine-readable database]. Technical report, Minneapolis: University of Minnesota.

Sojourner, A., West, K., and Mykerezi, M. (2012). Teacher performance pay can work: Evidence from adoption of Q-Comp in Minnesota. Unpublished manuscript. University of Minnesota.

Springer, M., Hamilton, L., McCaffrey, D., Ballou, D., Le, V., Pepper, M., Lockwood, J., and Stecher, B. (2010). Teacher pay for performance: Experimental evidence from the project on incentives in teaching. National Center on Performance Incentives.

Taylor, E. and Tyler, J. (2011). The effect of evaluation on performance: Evidence from longitudinal student achievement data of mid-career teachers. Technical report, National Bureau of Economic Research Working Paper No. 16877.

Temin, P. (2003). Low pay, low quality. *Education Next*, 3(3):8–13.

Tyler, J., Taylor, E., Kane, T., and Wooten, A. (2010). Using student performance data to identify effective classroom practices. *American Economic Review*, 100(2):256–60.

Vigdor, J. (2008). Scrap the sacrosanct salary schedule: How about more pay for new teachers, less for older ones? *Education Next*, 8.

West, K. and Mykerezi, E. (2011). Teachers' unions and compensation: The impact of collective bargaining on salary schedules and performance pay schemes. *Economics of Education Review*, 30(1):99–108.

Chapter 7

Appendix

7.1 Appendix to Chapter 2

7.1.1 Model Details

The timing is of the model such that the district and the teacher first agree to a contract. The teacher learns ρ_i for all i and executes the assigned lesson (if the contract is "direct") or selects and executes the lesson (if the contract is "delegate"). The teacher chooses effort, e_i. The district measures e_i (if the contract is "input") or y_i (if the contract is "output") and pays the teacher. I solve the problem using backward induction.

The district seeks to maximize student achievement net costs. Equation (7.1) describes the district's problem.

$$Max(E[y_i - w_i - m]) \tag{7.1}$$

The maximization is in expectation because the district does not know ρ_i *ex ante* when it agrees to a contract (recall that y_i is a function of ρ_i). Costs depend on wages, w_i, and monitoring costs, m. Wages are subscripted by i because the district can pay a different wage for each lesson. The district will pay a higher wage for a lesson that requires more effort on the teacher's behalf. The district wants to select a contract that solves (7.1) and does so by comparing the expected surplus from each contract option assuming that the teacher will respond rationally.

A rational teacher seeks to maximize utility. Equation (7.2) describes the teacher's problem.

$$Max(w_i - C(e_i) + B) \tag{7.2}$$

For simplicity, assume that the teacher is risk neutral.[1] Utility is a simple function of wages w_i, the cost of effort, $C(e_i)$, and any non-pecuniary benefit the teacher derives from teaching a given lesson, B. Let the teacher prefer one lesson deriving utility $B > 0$ from it. For all other lessons $B = 0$. Another way to characterize this would be to let every lesson have its own cost function with $C(e_j) < C(e_i)$ for some lesson j. In equation (7.2) I let one lesson cost $C(e) - B$ and every other lesson cost $C(e)$.

The wage-effort schedule is constant across lessons and the reservation wage is normalized to zero so momentarily ignoring the teacher's non-pecuniary benefit, the district should set $w_i - C(e_i^* = 0 \forall i$ where e_i^* is the optimal level of effort for lesson i.[2] Since the wage-effort schedule is constant across lessons, B need only to be positive to tip the teacher towards her preferred lesson. I assume that B is positive but small relative to the benefits the district derives from the teacher choosing the lesson that best fits a particular class and that the district knows the magnitude of B but not which lesson it is attached to.

Contract type: Input - Delegate

If the district delegates the choice of lesson to the teacher and pays based on inputs, the expected output is $E[y_i] = (\Sigma_{i=1}^{n}(\bar{\rho}_i + e_i^*))/n$. This is because the district assumes that each lesson will have average efficacy, i.e. $\rho_i = \bar{\rho}_i$, since it does not know how well each lesson fits the particular teacher or class. The expected output is the average of all possible lesson choices because the district assumes that the efficacy of lessons is uniformly distributed.

The district does not know *ex ante* which lesson the teacher will select but it does know that the teacher will choose her preferred lesson and thus the wage offered is the average

[1] This is simply so that I can abstract from the usual tradeoff between risk and incentives. The conclusion of the model is not dependent on this assumption.

[2] This effort level is "optimal" for both the district and the teacher. The district sets the wage offer such that the teacher chooses the level of effort that maximizes her utility and solves the district's maximization problem. Indeed this is the goal of a P4P contract – to design a pay scheme such that the agent is motivated to behave as if she were the principal.

cost of effort reduced by B, that is $w_i = \Sigma_{i=1}^{n} C(e_i^*) - B$. The district also incurs monitoring costs, m_e, so the resulting surplus is described by equation (7.3).

$$E[y_i - w_i - m] = (\Sigma_{i=1}^{n}(\bar{\rho}_i + e_i^*)/n) - (\Sigma_{i=1}^{n} C(e_i^*) - B) - m_e \tag{7.3}$$

Contract type: Input - Direct

Let lesson k be the lesson with the highest mean output. If the district directs the teacher to teach lesson k and pays based on inputs, the expected output is, $E[y_i] = \bar{\rho}_k + e_k^*$. In this case, the teacher has a $1/n$ chance of being assigned to teach her preferred lesson so the wage offered is the cost of effort on lesson k reduced by B/n so $w_i = C(e_k^*) - B/n$. Monitoring costs are again m_e and the resulting surplus is described by equation (7.4).

$$E[y_i - w_i - m] = (\bar{\rho}_k + e_k^*) - (C(e_k^*) - B/n) - m_e \tag{7.4}$$

Equation (7.4) > equation (7.3) so long as B is small relative to the distance of $\bar{\rho}_k$ from all other $\bar{\rho}_i$. This means that if the district is going to pay based on inputs, it should direct the choice of lesson unless it believes that all lessons are, on average, about the same in which case it can let the teacher choose the lesson and pay a slightly lower wage to capture some of teacher's rents.

Contract type: Output - Direct

If the district directs the choice of lesson and pays based on output, the resulting surplus is exactly the same as the input-direct contract with m_y in place of m_e. Expected surplus is described by equation (7.5).

$$E[y_i - w_i - m] = \bar{\rho}_k + e_k^* - (C(e_k^*) - B/n) - m_y \tag{7.5}$$

Equation (7.4) > equation (7.5) because of the assumption that $m_y > m_e$.

Contract type: Output - Delegate

At this point the input-direct contract dominates the other options. However, as long as the benefit the teacher derives from her preferred lesson is small relative to the gains in student

achievement from the optimal lesson, an output-delegate contract provides incentive for the teacher to use her knowledge about which lesson is best. Therefore, *ex ante* the district can be assured that the teacher will choose the optimal lesson. In what follows I show the necessary assumptions for an output-delegate contract to dominate the other options.

For sake of argument, let $\widehat{\rho_j} > \widehat{\rho_k}$, i.e. the district would have chosen the wrong lesson. The expected output is $E[y_i] = \rho_j + e_j^*$ and the wage offered is $w_i = C(e_j^*) - B/n$ because there is a $1/n$ chance that the teacher's preferred lesson will be lesson j. Measurement costs are m_y and equation (7.6) describes the expected surplus.

$$E[y_i - w_i - m] = \rho_j + e_j^* - (C(e_j^*) - B/n) - m_y \tag{7.6}$$

To see if equation (7.6) > equation (7.4), we need to know how much greater $\widehat{\rho_j}$ is than $\widehat{\rho_k}$. That is, in order to compare this contract to the others we have to quantify the benefit of the teacher's local knowledge and compare this to the cost of basing wages on output. Equation (7.6) > equation (7.4) if the benefit from using output based pay to extract the teacher's local knowledge exceeds the cost of using output based pay, i.e. costs such as lost productivity from teaching to the test.

Consider the following example. Assume that there is no difference in means, all the uncertainty is characterized by the variance. The lessons look identical to the district in expectation. Let both lessons be drawn from $\rho_i \sim N(0, \sigma^2)$. Since both are drawn from the same distribution we can compare them using order statistics where the first-order statistic is the minimum (i.e. the worst lesson) and the second-order statistic is the maximum (i.e. the best lesson). Denote the expectation of the second-order statistic as $E[\rho'_{2\{2\}}]$.

If the district does not know which lesson the teacher will choose, it expects that output will be $(\Sigma_{i=1}^{n}\bar{\rho_i} + e_i^*)/n = (\Sigma_{i=1}^{2}\bar{\rho_i} + e_i^*)/2 = 0 + e_i^*$. If the district knows the teacher will choose the best lesson it expects output will be $E[\rho'_{2\{2\}}] + e_j^* = \sigma/\sqrt{\pi} + e_j^*$ since $\sigma/\sqrt{\pi}$ is the expected value of the second order statistic of two draws from a normal distribution with mean 0 and variance σ^2. Therefore rewrite equation (7.6) as:

$$E[y_i - w_i - m] = \sigma/\sqrt{\pi} + e_j^* - C(e_j^*) + B/n - m_y \tag{7.7}$$

noting that $\bar{\rho_k} = 0$ in this example, (7.7) > (7.4) reduces to:

$$\sigma/\sqrt{\pi} > m_y - m_e \tag{7.8}$$

The inequality (7.8) clearly shows that the choice of contract depends on the relative costs of the two measurement options and the amount of uncertainty. Delegating and paying for output is preferred when inequality (7.8) holds. This is likely if the left hand side is big (lots of uncertainty) and/or the right hand side is small (few costs to output based monitoring). Since we have assumed that m_y includes the "costs" of multitasking, what equation (7.8) makes clear is that there is a tradeoff between making good use of the teacher's local knowledge and distorting incentives by focusing too much on testable outcomes.

7.1.2 Q-Comp Program Review Details

The MDE conducted a review of each Q-Comp program in 2009. As part of the review, they scored each district on a rubric. I construct a measure of the percent "proficient" or "exemplary" on each section of the rubric.

The job-embedded professional development (JEPD) section of the rubric has five subsections. These are:

1. Teachers can clearly describe the purpose and desired outcomes of their team meetings.
2. Team size and composition allow for professional development to be effectively delivered.
3. There is dedicated time for learning teams to meet weekly or every two weeks.
4. The teacher learning from the team meetings applies directly to classroom instruction.
5. The teacher learning from the team meetings has a connection to teacher observations.

A district that received scores of "proficient" or "exemplary" for four out of the five of these (but "below proficient" for one of the five) would have 80% proficient ($JEPDScore = 0.80$). Districts that have a higher percent proficient have implemented JEPD with more fidelity.

The evaluation section of the rubric has seven sub-sections. These are:

1. Teachers are observed multiple times a year by multiple trained observers.
2. A standard rubric is used.
3. All teachers are evaluated.
4. Observers receive initial training.
5. Observers receive ongoing training.
6. Teachers receive training regarding the rubric.
7. Pre and post evaluation conferences promote reflection.

A district that received scores of "proficient" or "exemplary" for four out of the seven of these (but "below proficient" for three of the seven) would have 57% proficient ($EvalScore = 0.57$). Districts that have a higher percent proficient have implemented the evaluation process with more fidelity.

7.1.3 Additional Results

Table 7.1: The Effect of Q-Comp - Displaying Demographic Covariates

Outcome test:	MCA		NWEA	
Subject:	Reading	Math	Reading	Math
1(post-adoption)	0.031**	0.004	0.032**	0.038
	(0.015)	(0.021)	(0.016)	(0.026)
Share male	-0.083***	-0.134***	-0.057	-0.121**
	(0.029)	(0.038)	(0.052)	(0.061)
Share free lunch	0.104***	0.104***	0.029	0.045
	(0.036)	(0.034)	(0.047)	(0.066)
Share special education	-0.046	-0.209***	0.132*	0.097
	(0.040)	(0.071)	(0.067)	(0.080)
Share African-American	-0.143***	-0.095***	-0.001	-0.154
	(0.030)	(0.030)	(0.088)	(0.113)
Share Hispanic	-0.075*	-0.007	-0.102	-0.084
	(0.043)	(0.045)	(0.108)	(0.109)
Share Asian-American	0.021	0.068	0.086	0.009
	(0.046)	(0.043)	(0.149)	(0.171)
Share Native American	-0.131***	-0.050	0.028	0.016
	(0.050)	(0.052)	(0.063)	(0.093)
Enrollment	-0.004	-0.011	0.005	-0.010
	(0.006)	(0.008)	(0.008)	(0.009)
Districts	369	369	273	273
Students	696,970	686,483	247,026	247,767
Student-years	2,052,337	2,007,029	651,891	655,341
Adj. R^2	0.774	0.793	0.793	0.838

Dependent variable is student levels score on the MCA and NWEA Assessments normalized to mean 0 and standard deviation 1. Standard errors are corrected for heteroskedasticity and correlation within district. Significance *** 1%, ** 5%, * 10%. All specifications include year indicators, student fixed effects, and indicator for having dropped Q-Comp and covariates that measure school-grade-year student demographics. NWEA models also include a covariate that measures days elapsed between tests.

Table 7.2: The effect of Q-Comp - Models with All Interaction Terms

Outcome test:	MCA		NWEA	
Subject:	Reading	Math	Reading	Math
1(post-adoption)	-.198	0.003	0.105	0.157
	(0.164)	(0.167)	(0.123)	(0.255)
Teacher P4P$	0.014	-.080	-.120	-.085
	(0.128)	(0.124)	(0.081)	(0.159)
School P4P$	0.063	0.177	-.141	-.430
	(0.211)	(0.244)	(0.172)	(0.381)
Evaluation P4P$	0.167*	0.095	0.005	0.22
	(0.088)	(0.093)	(0.072)	(0.149)
JEPD Score	-.035	-.430	-.545***	-.210
	(0.313)	(0.346)	(0.173)	(0.416)
Eval Score	0.28	0.412*	0.568***	-.104
	(0.183)	(0.244)	(0.128)	(0.292)
(Teacher P4P$)*(JEPD Score)	0.052	0.068	0.415***	0.194
	(0.217)	(0.208)	(0.103)	(0.226)
(Teacher P4P$)*(Eval Score)	-.057	0.053	-.304***	-.079
	(0.103)	(0.12)	(0.083)	(0.177)
(School P4P$)*(JEPD Score)	0.082	-.044	-.170	-.574
	(0.379)	(0.412)	(0.185)	(0.405)
(School P4P$)*(Eval Score)	-.134	-.166	0.308	1.286**
	(0.351)	(0.348)	(0.255)	(0.545)
(Evaluation P4P$)*(JEPD Score)	-.043	0.221	0.229**	-.047
	(0.126)	(0.191)	(0.104)	(0.24)
(Evaluation P4P$)*(Eval Score)	-.141	-.344**	-.292***	-.120
	(0.1)	(0.134)	(0.069)	(0.165)
N Student-years	2,052,337	2,007,029	651,891	655,341
N Students	696,970	686,484	247,026	247,768
N Districts	369	369	273	273
Adjusted R^2	0.774	0.793	0.793	0.838

Dependent variable is student levels score on the MCA and NWEA Assessments normalized to mean 0 and standard deviation 1. Standard errors are corrected for heteroskedasticity and correlation within district. Significance *** 1%, ** 5%, * 10%. All specifications include year indicators, student fixed effects, and indicator for having dropped Q-Comp and covariates that measure school-grade-year student demographics. NWEA models also include a covariate that measures days elapsed between tests.

Table 7.3: The Effect of Q-Comp on MCA - Models with Pre-adoption Terms

	Reading	Reading	Math	Math
1(post-adoption)	0.028**	-.124	-.004	-.261**
	(0.012)	(0.111)	(0.017)	(0.129)
Teacher P4P$		-.002		0.046
		(0.101)		(0.123)
School P4P$		0.023		0.012
		(0.037)		(0.046)
Evaluation P4P$		0.129**		0.251***
		(0.055)		(0.063)
JEPD Score		-.052		-.072
		(0.104)		(0.123)
Eval Score		0.195**		0.365***
		(0.088)		(0.125)
(Teacher P4P$)*(JEPD Score)		0.025		-.040
		(0.134)		(0.165)
(Evaluation P4P$)*(Eval Score)		-.138*		-.307***
		(0.072)		(0.09)
1(2+ yrs. pre-QComp)	-.010	0.009	-.026	-.039*
	(0.018)	(0.023)	(0.019)	(0.023)
1(2+ yrs pre-Qcomp)*(Teacher P4P$)		0.002		0.0005
		(0.02)		(0.023)
1(2+ yrs pre-Qcomp)*(School P4P$)		-.053		-.002
		(0.055)		(0.061)
1(2+ yrs pre-Qcomp)*(Evaluation P4P$)		-.008		0.013
		(0.02)		(0.022)
N Student-years	2,052,337	2,052,337	2,007,029	2,007,029
N Students	696,969	696,969	686,483	686,483
N Districts	369	369	369	369
Adjusted R^2	0.774	0.774	0.792	0.793

Dependent variable is student levels score on the MCA Reading and Math Assessments normalized to mean 0 and standard deviation 1. Standard errors are corrected for heteroskedasticity and correlation within district. Significance *** 1%, ** 5%, * 10%. All specifications include year indicators, student fixed effects, and indicator for having dropped Q-Comp and covariates that measure school-grade-year student demographics.

Table 7.4: The Effect of Q-Comp on NWEA - Models with Pre-adoption Terms

	Reading	Reading	Math	Math
1(post-adoption)	0.032**	-.022	0.036*	0.293
	(0.012)	(0.083)	(0.02)	(0.182)
Teacher P4P$		-.088		-.293**
		(0.07)		(0.147)
School P4P$		-.039		0.025
		(0.043)		(0.056)
Evaluation P4P$		0.078*		0.102
		(0.04)		(0.086)
JEPD Score		-.063		-.470***
		(0.082)		(0.16)
Eval Score		0.141*		0.055
		(0.085)		(0.175)
(Teacher P4P$)*(JEPD Score)		0.135		0.359**
		(0.086)		(0.177)
(Evaluation P4P$)*(Eval Score)		-.104*		-.056
		(0.054)		(0.128)
1(2+ yrs. pre-QComp)	0.001	0.015	-.011	0.026
	(0.017)	(0.021)	(0.028)	(0.035)
1(2+ yrs pre-Qcomp)*(Teacher P4P$)		-.042		-.057
		(0.031)		(0.065)
1(2+ yrs pre-Qcomp)*(School P4P$)		0.113		0.139
		(0.081)		(0.11)
1(2+ yrs pre-Qcomp)*(Evaluation P4P$)		-.013		-.035
		(0.015)		(0.03)
N Student-years	651,891	651,891	655,341	655,341
N Students	247,025	247,025	247,767	247,767
N Districts	273	273	273	273
Adjusted R^2	0.793	0.793	0.838	0.838

Dependent variable is student levels score on the NWEA Reading and Math Assessments normalized to mean 0 and standard deviation 1. Standard errors are corrected for heteroskedasticity and correlation within district. Significance *** 1%, ** 5%, * 10%. All specifications include year indicators, student fixed effects, and indicator for having dropped Q-Comp and covariates that measure school-grade-year student demographics and days elapsed between tests.

Table 7.5: The Effect of Q-Comp - Models with Transformed Interaction Terms

Outcome test:	MCA		NWEA	
Subject:	Reading	Math	Reading	Math
1(post-adoption)	-.040	-.024	-.079	0.141
	(0.08)	(0.104)	(0.087)	(0.161)
Teacher P4P$	0.029	0.002	0.038	-.051
	(0.039)	(0.043)	(0.039)	(0.084)
School P4P$	0.037	-.008	-.063	-.004
	(0.045)	(0.049)	(0.049)	(0.063)
Evaluation P4P$	0.043*	0.075***	0.039	0.104***
	(0.023)	(0.023)	(0.027)	(0.032)
JEPD Score	-.033	-.144	0.043	-.272*
	(0.083)	(0.089)	(0.069)	(0.142)
Eval Score	0.042	0.119	0.077	0.052
	(0.067)	(0.097)	(0.085)	(0.091)
(Teacher P4P$)*(JEPD Score 0-1)	-.010	0.02	-.011	0.039
	(0.026)	(0.03)	(0.026)	(0.058)
(Evaluation P4P$)*(Eval Score 0-1)	-.010	-.062**	-.038	-.052
	(0.022)	(0.024)	(0.027)	(0.043)
N Student-years	2052337	2007029	651891	655341
N Students	696969	686483	247025	247767
N Districts	369	369	273	273
Adjusted R^2	0.774	0.793	0.793	0.838

Dependent variable is student levels score on the MCA and NWEA Assessments normalized to mean 0 and standard deviation 1. Standard errors are corrected for heteroskedasticity and correlation within district. Significance *** 1%, ** 5%, * 10%. All specifications include year indicators, student fixed effects, and indicator for having dropped Q-Comp and covariates that measure school-grade-year student demographics. NWEA models also include a covariate that measures days elapsed between tests.

7.2 Appendix to Chapter 3

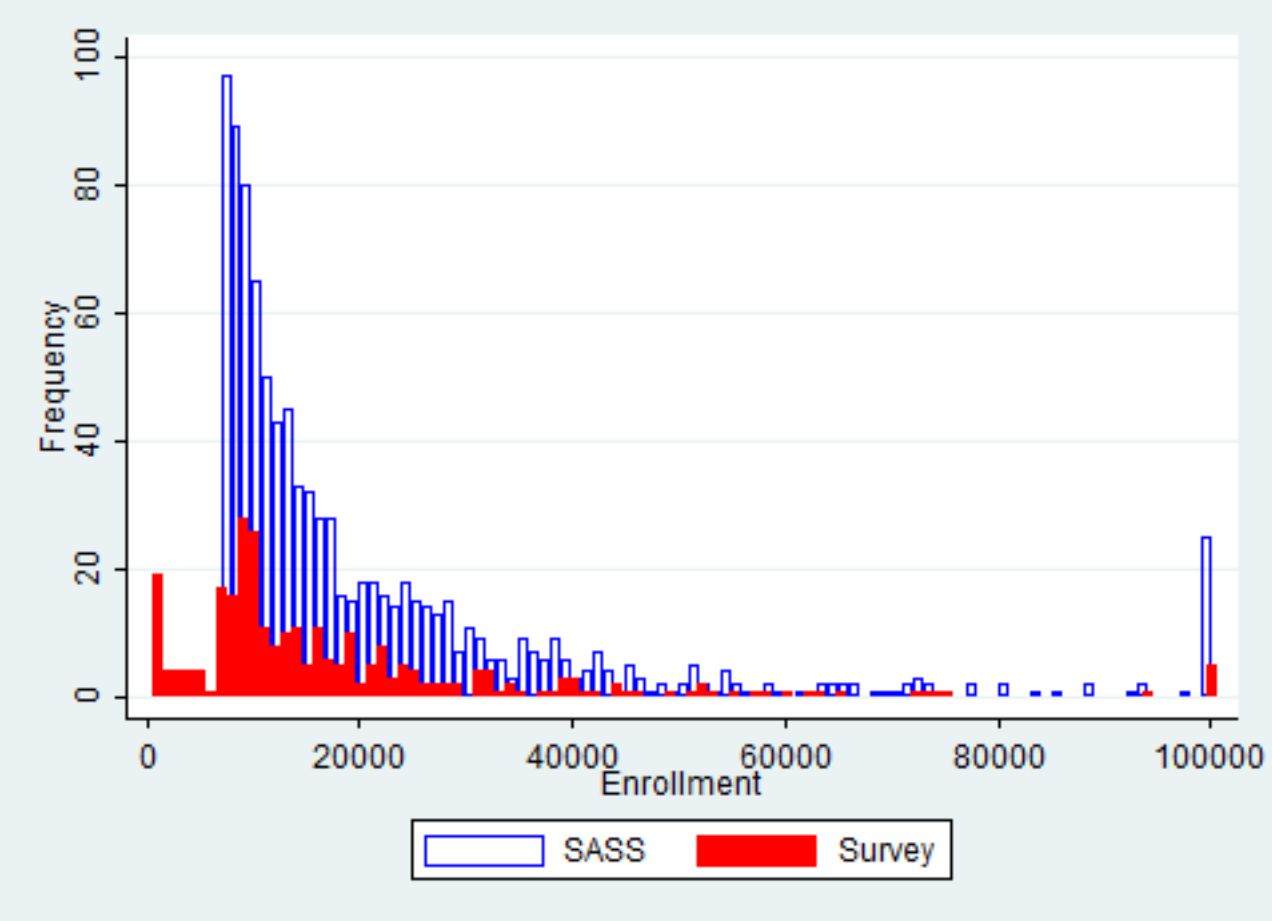
7.2.1 Survey Analysis

Table 7.6: Districts Surveyed by State

State	Freq.	Share	State	Freq.	Share
AK	1	0.24	MT	1	0.24
AL	4	0.98	NC	10	2.44
AR	4	0.98	ND	1	0.24
AZ	5	1.22	NE	4	0.98
CA	72	17.56	NJ	3	0.73
CO	6	1.46	NM	2	0.49
CT	1	0.24	NV	4	0.98
DE	1	0.24	NY	8	1.95
FL	11	2.68	OH	7	1.71
GA	9	2.20	OK	4	0.98
IA	4	0.98	OR	7	1.71
ID	2	0.49	PA	9	2.20
IL	13	3.17	SC	4	0.98
IN	9	2.20	SD	1	0.24
KS	3	0.73	TN	6	1.46
KY	1	0.24	TX	26	6.34
LA	6	1.46	UT	3	0.73
MA	3	0.73	VA	13	3.17
MD	5	1.22	WA	10	2.44
MI	13	3.17	WI	11	2.68
MN	92	22.44	WV	1	0.24
MO	6	1.46	WY	1	0.24
MS	3	0.73			

This table lists the breakdown of districts that responded to my survey by state. Districts in Minnesota were oversampled. This was done to assist with a concurrent project that focuses on P4P in Minnesota. A complete list of the districts that responded to the survey is available upon request.

Figure 7.1: Comparison of Districts Surveyed and a Nationally Representative Sample (SASS)



Of the 410 districts that completed my survey, 282 were also sampled in the 2007-08 SASS. This figure compares enrollments in these districts to all districts in the 2007-08 SASS with at least 7,000 students. Note that, although not shown on this figure, the SASS surveys districts with fewer than 7,000 students. Indeed, they are not shown because the frequency exceeds 100. Figure7.1 shows that my sample is smaller than, but consistent with, the SASS. This supports the assertion that my survey is a representative sample of districts with at least 7,000 students.

Table 7.7: Probit Estimate of Survey Completion

Total students	7.48e-06
	(8.57e-06)
Full time equivalent staff	0.00005
	(0.0001)
Number of schools	-.008*
	(0.004)
Pupil-teacher ratio	0.017
	(0.03)
Share pop college educ	-.189
	(1.274)
Per capita income	-1.00e-05
	(0.00002)
Share pop below poverty	0.997
	(1.044)
Share male students	-.837
	(3.992)
Share black students	0.277
	(0.595)
Share white students	0.603
	(0.524)
Share hispanic students	0.286
	(0.531)
Share FRL students	-.428
	(0.332)
Share ELL students	0.303
	(0.62)
Share special education	0.886
	(1.871)
N	1393
$PsuedoR^2$	0.141

Dependent variable = 1 if survey completed. Table reports results for probit estimation. State fixed effects and an indicator for charter school status are also included but not reported here. Significance: $*$: 10% $**$: 5% $***$: 1%.

7.2.2 Instrumental Variables Estimation

As stated in the section entitled "Robustness and alternative specifications," all results presented in the body of the paper are from ordinary least squares (OLS) and probit models. It is possible, however, that these models are not ideal because they do not deal with potential endogeneity of unions and pay and/or tenure policies. Here I present results using the instrument described in the text. Note that since all the instruments are state level variables I can not include state fixed effects, therefore, the Table 7.8 reports results analogous to Table 3.4 column (1c) in the main text, i.e. those will demographic controls and region fixed effects.

The various instruments perform well on most diagnostic tests. Specifically, first stage results show that the instruments are all strong predictors of collective bargaining. None of the instruments appear to be weak, yielding test statistics that easily surpass the Stock and Yogo critical values. Additionally, reduced form estimates (OLS regressions of pay on the instruments and exogenous variables only) indicate that the instruments have the expected association with pay (all are positively associated with teacher pay). The only concern is that a test of overidentification restrictions based on Sargan's statistic rejects the null hypotheses of non-correlation of the instruments with the errors. This casts doubt on the exogeneity of the instruments.

Table 7.8: The Effect of Collective Bargaining – Instrumental Variable Results

	OLS	Instrumental variables			
Dependent variable		State law (historic)	State law (current)	Share unionized in 1964	Share immigrants from 1900-1930
Starting salary	0.0626***	0.184***	0.158***	0.510***	0.298***
	(0.0131)	(0.0256)	(0.0158)	(0.0443)	(0.0440)
Returns to experience	0.00502***	0.00445**	-0.00224	0.0301***	0.0355***
	(0.000715)	(0.00214)	(0.00184)	(0.00315)	(0.00484)
Returns to degree	0.0187***	0.0292***	0.00924	0.0457***	0.00448
	(0.00449)	(0.00738)	(0.00879)	(0.0137)	(0.0192)
Salary for BA + 10 yrs	0.0923***	0.211***	0.153***	0.758***	0.596***
	(0.0118)	(0.0392)	(0.0182)	(0.0578)	(0.0643)
Salary for MA + 10 yrs	0.105***	0.229***	0.158***	0.813***	0.591***
	(0.0109)	(0.0390)	(0.0166)	(0.0621)	(0.0613)
Reward shortage field	0.00280	-0.213***	-0.198***	-0.429***	0.188**
	(0.0135)	(0.0274)	(0.0337)	(0.0650)	(0.0786)
Reward national board certification	0.0706***	0.145***	0.0569	-0.0447	-0.465***
	(0.0229)	(0.0374)	(0.0514)	(0.0607)	(0.0948)
Reward excellence in teaching	-0.00247	-0.00249	-0.0423	-0.228***	-0.0561
	(0.0105)	(0.0265)	(0.0303)	(0.0339)	(0.0639)
Senior teachers dismissed	-0.00924	0.0313**	0.0171	0.0173	0.00450
	(0.0101)	(0.0154)	(0.0170)	(0.0367)	(0.0489)
Junior teachers dismissed	-0.00238	0.00245	0.00626	0.0128	-0.000377
	(0.00539)	(0.00955)	(0.00940)	(0.0172)	(0.0216)

Table reports β_1 from OLS and IV estimations of equation (3.1) with various dependent variables. Robust standard errors are in parentheses. Significance: $*$: 10% $**$: 5% $***$: 1%. Control variables include total enrollment, racial composition of the student body, the share of students that receive free and reduced lunch, the total number of teachers, racial composition of the teachers, the share of school aged children in the district, the share of college educated residents in the district, per capita income and census region fixed effects.

7.2.3 District Random Effects

Table 7.9: Summary of Collective Bargaining Status for Districts in Both Waves of the SASS

	Collective Bargaining 2007-08		
2003-04	No	Yes	Total
No	827	74	901
Yes	161	1,153	1,314
Total	988	1,227	2,215

Table 7.10: The Effect of Collective Bargaining – Random Effects Results

Dependent variable	OLS	Random Effects
Starting salary	0.0682***	0.0698***
	(0.00294)	(0.00305)
Returns to experience	0.00449***	0.00466***
	(0.000311)	(0.000312)
Returns to degree	0.0218***	0.0202***
	(0.00174)	(0.00180)
Salary for BA + 10 yrs	0.0912***	0.0942***
	(0.00345)	(0.00356)
Salary for MA + 10 yrs	0.107***	0.107***
	(0.00355)	(0.00366)

Table reports β_1 from estimates of equation (3.1) with various dependent variables. The first column reports results from OLS regressions. These are analogous to Table 3.4 column (1c) but without proper sampling weights. The second column report results from a random effect maximum likelihood regression. Standard errors are in parentheses. Significance: $*$: 10% $**$: 5% $***$: 1%. Control variables include total enrollment, racial composition of the student body, the share of students that receive free and reduced lunch, the total number of teachers, racial composition of the teachers, the share of school aged children in the district, the share of college educated residents in the district, per capita income and census region fixed effects.

7.3 Appendix to Chapter 4

Table 7.11: The Teacher Wage Gap, Results for Hourly Wages Using Usual and Diary Hours of work – Female Respondents only

	(1) Usual Hours	(2) Diary Hours	(3) Usual Hours	(4) Diary Hours
Teacher	-.109***	0.054***		
	(0.018)	(0.018)		
Elementary/middle school			-.116***	0.06***
			(0.022)	(0.021)
Secondary school			-.134***	-.024
			(0.028)	(0.025)
Special education			-.009	0.171***
			(0.046)	(0.045)
Age	0.06***	0.062***	0.059***	0.062***
	(0.006)	(0.006)	(0.006)	(0.006)
Age^2	-.0006***	-.0006***	-.0006***	-.0006***
	(0.00008)	(0.00007)	(0.00008)	(0.00007)
Masters degree	0.188***	0.22***	0.187***	0.218***
	(0.016)	(0.017)	(0.016)	(0.017)
PhD degree	0.28***	0.312***	0.279***	0.311***
	(0.052)	(0.05)	(0.051)	(0.05)
Professional degree	0.361***	0.385***	0.361***	0.385***
	(0.049)	(0.051)	(0.049)	(0.051)
White	-.005	0.01	-.005	0.01
	(0.018)	(0.018)	(0.018)	(0.018)
N	8943	8943	8943	8943
R^2	0.096	0.111	0.097	0.112

This table is equivalent to Table 4.3 in the text except it includes only female respondents. The dependent variable is ln(hourly wage) calculated from equation (4.4) using usual hours of work or equation (4.5) using diary hours of work. Significance *** 1%, ** 5%, * 10%. Each observation is weighted using the weights provided by the ATUS and standard errors are calculated with successive difference replication (SDR) variance estimation using the replicate weights provided by the ATUS. Sample includes female ATUS respondents with at least a Bachelors degree who are full time workers with positive weekly earnings.

Table 7.12: The Teacher Wage Gap, Results for Hourly Wages Using Usual and Diary Hours of Work – Male Respondents Only

	(1) Usual Hours	(2) Diary Hours	(3) Usual Hours	(4) Diary Hours
Teacher	-.240***	-.130***		
	(0.029)	(0.03)		
Elementary/middle school			-.196***	-.067*
			(0.038)	(0.037)
Secondary school			-.288***	-.199***
			(0.037)	(0.036)
Special education			-.114	0.055
			(0.13)	(0.143)
Age	0.066***	0.074***	0.065***	0.074***
	(0.005)	(0.005)	(0.005)	(0.005)
Age^2	-.0007***	-.0007***	-.0007***	-.0007***
	(0.00006)	(0.00006)	(0.00006)	(0.00006)
Masters degree	0.117***	0.126***	0.118***	0.126***
	(0.021)	(0.023)	(0.021)	(0.023)
PhD degree	0.175***	0.225***	0.175***	0.226***
	(0.027)	(0.027)	(0.027)	(0.027)
Professional degree	0.283***	0.33***	0.285***	0.332***
	(0.028)	(0.027)	(0.028)	(0.027)
White	0.041*	0.076***	0.042**	0.077***
	(0.021)	(0.022)	(0.021)	(0.022)
N	9211	9211	9211	9211
R^2	0.092	0.099	0.093	0.1

This table is equivalent to Table 4.3 in the text except it includes only male respondents. The dependent variable is ln(hourly wage) calculated from equation (4.4) using usual hours of work or equation (4.5) using diary hours of work. Significance *** 1%, ** 5%, * 10%. Each observation is weighted using the weights provided by the ATUS and standard errors are calculated with successive difference replication (SDR) variance estimation using the replicate weights provided by the ATUS. Sample includes male ATUS respondents with at least a Bachelors degree who are full time workers with positive weekly earnings.

Table 7.13: The Teacher Wage Gap, Results for Hourly Wages Using Usual and Diary Hours of Work – Respondents Whose Highest Degree is a Bachelors

	(1) Usual Hours	(2) Diary Hours	(3) Usual Hours	(4) Diary Hours
Teacher	-.158***	0.01		
	(0.02)	(0.02)		
Elementary/middle school			-.158***	0.031
			(0.027)	(0.025)
Secondary school			-.188***	-.070***
			(0.028)	(0.026)
Special education			-.041	0.145*
			(0.069)	(0.081)
Age	0.067***	0.073***	0.067***	0.073***
	(0.005)	(0.005)	(0.005)	(0.005)
Age^2	-.0007***	-.0008***	-.0007***	-.0008***
	(0.00006)	(0.00006)	(0.00006)	(0.00006)
White	0.041**	0.052***	0.041**	0.052***
	(0.016)	(0.017)	(0.016)	(0.017)
Female	-.182***	-.214***	-.183***	-.216***
	(0.014)	(0.015)	(0.014)	(0.015)
N	11392	11392	11392	11392
R^2	0.111	0.112	0.112	0.113

This table is equivalent to Table 4.3 in the text except it includes only respondents whose highest degree is a BA. The dependent variable is ln(hourly wage) calculated from equation (4.4) using usual hours of work or equation (4.5) using diary hours of work. Significance *** 1%, ** 5%, * 10%. Each observation is weighted using the weights provided by the ATUS and standard errors are calculated with successive difference replication (SDR) variance estimation using the replicate weights provided by the ATUS. Sample includes ATUS respondents whose highest degree is a BA who are full time workers with positive weekly earnings.

Table 7.14: The Teacher Wage Gap, Results for Hourly Wages Using Usual and Diary Hours of Work – Respondents Whose Highest Degree is a Masters

	(1) Usual Hours	(2) Diary Hours	(3) Usual Hours	(4) Diary Hours
Teacher	-.135***	-.012		
	(0.024)	(0.025)		
Elementary/middle school			-.104***	0.031
			(0.026)	(0.026)
Secondary school			-.242***	-.157***
			(0.043)	(0.043)
Special education			-.011	0.149***
			(0.05)	(0.045)
Age	0.046***	0.058***	0.045***	0.057***
	(0.008)	(0.008)	(0.008)	(0.008)
Age^2	-.0005***	-.0006***	-.0004***	-.0006***
	(0.00009)	(0.00009)	(0.00009)	(0.00009)
White	-.033	0.01	-.032	0.011
	(0.022)	(0.023)	(0.022)	(0.023)
Female	-.120***	-.114***	-.128***	-.125***
	(0.021)	(0.023)	(0.021)	(0.023)
N	4953	4953	4953	4953
R^2	0.045	0.04	0.048	0.046

This table is equivalent to Table 4.3 in the text except it includes only respondents whose highest degree is an MA. The dependent variable is ln(hourly wage) calculated from equation (4.4) using usual hours of work or equation (4.5) using diary hours of work. Significance *** 1%, ** 5%, * 10%. Each observation is weighted using the weights provided by the ATUS and standard errors are calculated with successive difference replication (SDR) variance estimation using the replicate weights provided by the ATUS. Sample includes ATUS respondents whose highest degree is an MA who are full time workers with positive weekly earnings.

Table 7.15: The Teacher Wage Gap, Results for Hourly Wages Using Usual and Diary Hours of Work – By Sector

	(1) Usual Hours	(2) Diary Hours	(3) Usual Hours	(4) Diary Hours
Public teacher	-.126***	0.019		
	(0.017)	(0.017)		
Private teacher	-.284***	-.107***		
	(0.04)	(0.037)		
Public elementary/middle school			-.112***	0.055***
			(0.019)	(0.019)
Private elementary/middle school			-.301***	-.120**
			(0.052)	(0.049)
Public secondary school			-.184***	-.098***
			(0.027)	(0.027)
Private secondary school			-.350***	-.173***
			(0.043)	(0.041)
Public special education			-.030	0.152***
			(0.046)	(0.046)
Private special education			-.007	0.162*
			(0.102)	(0.096)
Age	0.062***	0.068***	0.061***	0.068***
	(0.004)	(0.004)	(0.004)	(0.004)
Age^2	-.0006***	-.0007***	-.0006***	-.0007***
	(0.00005)	(0.00005)	(0.00005)	(0.00005)
Masters degree	0.151***	0.171***	0.15***	0.17***
	(0.014)	(0.015)	(0.014)	(0.015)
PhD degree	0.217***	0.261***	0.216***	0.26***
	(0.025)	(0.025)	(0.025)	(0.025)
Professional degree	0.317***	0.355***	0.318***	0.357***
	(0.027)	(0.026)	(0.027)	(0.026)
White	0.021	0.046***	0.021	0.047***
	(0.013)	(0.014)	(0.013)	(0.014)
Female	-.159***	-.183***	-.162***	-.187***
	(0.011)	(0.011)	(0.011)	(0.012)
N	18154	18154	18154	18154
R^2	0.12	0.129	0.121	0.131

This table is equivalent to Table 4.3 in the text except it splits teachers into those who work in the public sector and those who work in the private sector. The dependent variable is ln(hourly wage) calculated from equation (4.4) using usual hours of work or equation (4.5) using diary hours of work. Significance *** 1%, ** 5%, * 10%. Each observation is weighted using the weights provided by the ATUS and standard errors are calculated with successive difference replication (SDR) variance estimation using the replicate weights provided by the ATUS. Sample includes female ATUS respondents with at least a Bachelors degree who are full time workers with positive weekly earnings.