

COMPLEMENTARITY OF INCENTIVE PAY AND DECENTRALIZED DECISION MAKING:  
EVIDENCE FROM MINNESOTA'S Q-COMP PROGRAM FOR TEACHERS

Kristine L. West\*

November 25, 2011

**Abstract**

Since 2005, dozens of Minnesota school districts have implemented “pay for performance” (P4P) and a decentralized approach to professional development as part of the state’s Quality Compensation (Q-Comp) program. This paper uses the experience of Q-Comp to offer new evidence on the complementarity of output based P4P and decentralized decision making. Teachers are asymmetrically informed about what actions are best for their specific classes: they have local knowledge about their own abilities and preferences and their students’ abilities and preferences and the quality of the match between these. This paper describes a theoretical model that shows, given this environment, districts should offer output based P4P coupled with support for teachers to utilize their local knowledge. I exploit district-level variation in Q-Comp plans and the timing of adoption and find that districts which put more dollars at risk for teacher- or small team-level goals set via a decentralized professional development process experienced the largest gains in student achievement. P4P and decentralized professional development individually also lead to gains, but the evidence points to significant complementarities of the two reforms.

**JEL Classifications: M5, D8, J3, J4, I2**

\*U. Minnesota, Department of Applied Economics 337J Ruttan Hall, St. Paul MN 55105. Thanks to Joe Ritter for comments and to Kristie Anderson of the Minnesota Department of Education for help with data acquisition. Fellowship support from the Joseph M. Juran Center is gratefully acknowledged. All errors are my own.

# 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 tie teacher compensation to student outcomes, however, these “pay-for-performance” (P4P) reforms are 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 and very few rely on measures of teacher effectiveness such as “value-added” scores. 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.<sup>1</sup> 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.

Output based P4P contracts in education are 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

---

<sup>1</sup>“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.

erode cooperation amongst teachers. Linking teacher pay to student outcomes introduces dangerous incentives for cheating and other unproductive actions.

In this paper, I offer a theoretical model adapted 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 this type of contract.

This theoretical model takes into account education’s unique production technology. Specifically, teachers are asymmetrically informed about their classes and thus better placed than district officials to make choices about curriculum and pedagogy. The theoretical models that advocates and detractors of output based P4P in education more commonly cite do not incorporate the importance of teachers’ local knowledge about their own preferences and abilities, the preferences and abilities of their students and the quality of the match between these. 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 a management style 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.<sup>2</sup> Districts that use output based P4P contracts

---

<sup>2</sup>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

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. I find that districts that implement reforms intended to support teacher-level decision making about curriculum and pedagogy experience gains in student achievement and, most importantly, that output based P4P and support for teacher-level decision making are complements. That is, districts that implement both output based P4P and decentralized professional development experience the largest gains.

The paper proceeds as follows. Section 2 summarizes the theoretical model and builds intuition for how it relates to education. Section 3 provides evidence that education fits the key assumptions of the model. Section 4 discusses the unique features of Q-Comp and the data available which allow for an empirical test of the model’s predictions. Section 5 outlines the empirical strategy and section 6 presents the results. Section 7 concludes with a discussion of how these results relate to current educational policy debates.

---

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 who will have the right to decide what task is rewarded and how the agent will be held accountable.

## 2 Theory

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 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.<sup>3</sup>

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".

The 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 will provide 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.<sup>4</sup> The argument is that since teaching is a multi-dimensional

---

<sup>3</sup>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.

<sup>4</sup>In fact, Holmstrom and Milgrom (1991) use teachers and teaching to the test (or even cheating) as the motivating example in their seminal paper.

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

Consider the following illustration. 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,  $\mu_i$  and the variance,  $\sigma_i^2$ . The variance is a measure of how much technological uncertainty there is.<sup>5</sup> That is, for a given vector of means,  $\sigma^2$  measures how sure the district is about which lesson is best. A high  $\sigma^2$  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 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.<sup>6</sup>

In figure 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 III, there is a lot of technological uncertainty,  $\sigma^2 = 2$ . The district is very unsure and thus the teacher’s local knowledge is much more 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.

---

<sup>5</sup>Lessons are indexed by  $i$ . I assume the variance is constant across all lessons so in what follows I drop the subscript  $i$  on  $\sigma^2$ .

<sup>6</sup>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 I: Moderate Uncertainty

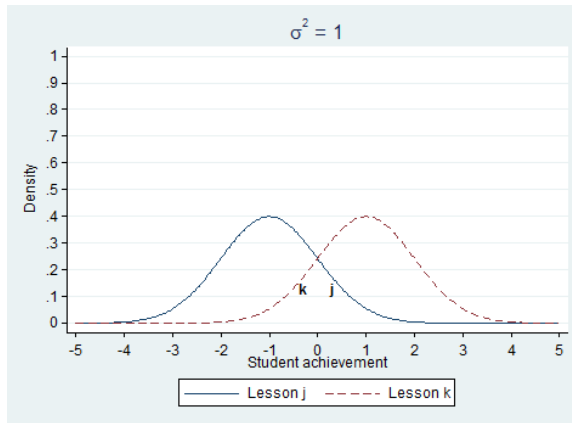


Figure II: Less uncertainty (variance)

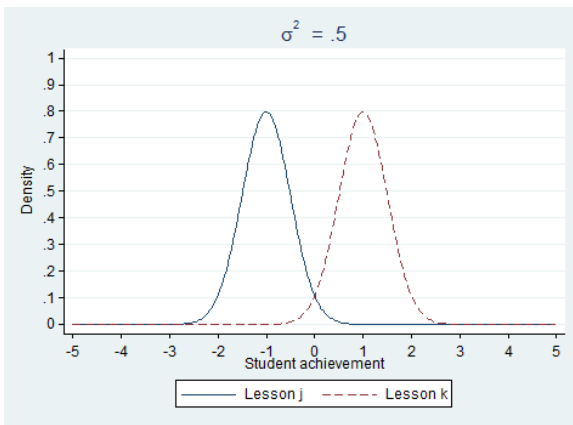


Figure III: More uncertainty (variance)

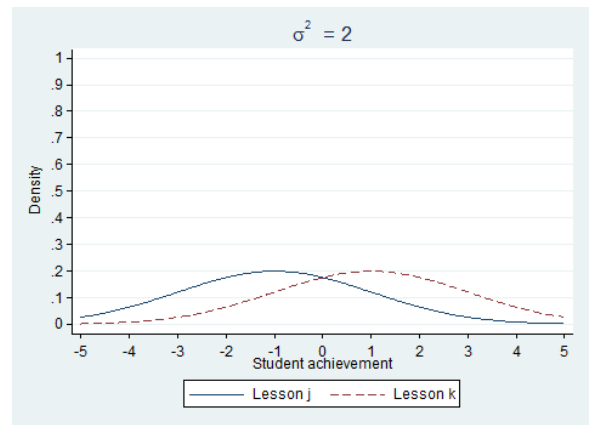


Figure IV: Less uncertainty (means)

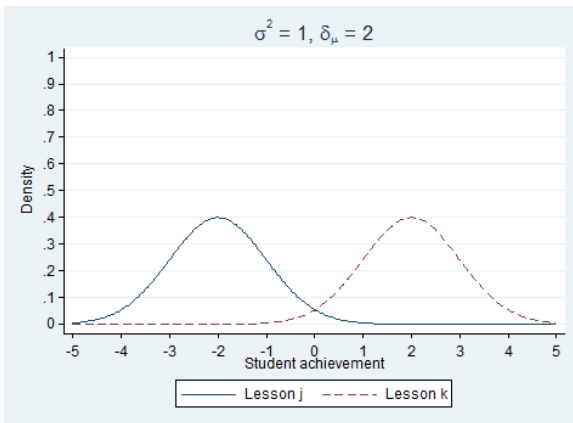
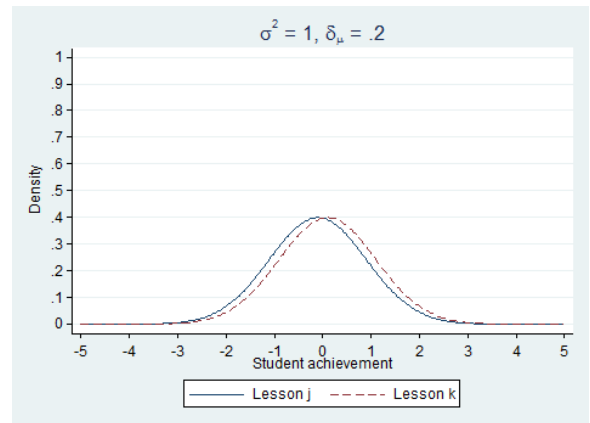


Figure V: More uncertainty (means)



Another way to characterize technological uncertainty is through the difference in the means. In this case, hold  $\sigma^2$  constant and allow the means to be relatively far apart, figure IV, or relatively close, figure 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.<sup>7</sup> The main result of the model is that if there is little technological uncertainty (figures II and IV) the district should direct the teacher to teach lesson k. If there is a lot of technological uncertainty (figures III and V) the district should delegate the choice lesson to the teacher.

Further, if I assume that in education it is costly to measure student outcomes because it may lead to unproductive actions such as “teaching to the test”, this cost must be weighed against the benefit of choosing the best lesson. If there is sufficient technological uncertainty, the benefit of choosing the best lesson outweighs the cost of “teaching to the test”. The district should pay based on inputs when it directs the choice of lesson but it should pay based on outputs when it delegates 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.

I describe the model in detail and derive these results in the appendix. In the following section I outline the basics of the model and its main conclusions to provide motivation for an empirical test using data from Minnesota’s Q-Comp program. Q-Comp provides a unique opportunity to test the model’s predictions and I find support for the conclusion that output based P4P and delegated decision making are complements.

---

<sup>7</sup>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.

## 2.1 Optimal Contract: The Role of Uncertainty

Consider a district (the principal) who hires a teacher (the agent) to exert effort on one of  $n$  possible lessons. The teacher exerts effort on lesson  $i$ ,  $e_i$ , 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 unobservable teacher/student/class idiosyncrasies,  $\rho_i$ .

$$y_i = \rho_i + e_i \tag{1}$$

From the point of view of the district,  $\rho_i$  are random. The district only knows the distributions of  $\rho_i$ . As described above, assume that all  $\rho_i$  have the same variance,  $\sigma^2$ , but differ in their means,  $\mu_i$ . 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, learns the true value,  $\hat{\rho}_i$ , 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 incentives for the teacher to chose the best lesson rather than her preferred lesson.

Assuming that the teacher does not necessarily prefer the lesson that 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 may 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  $m_y$ . These can be direct costs such as an administrator's time to conduct an evaluation. Paying based on student test scores may seem relatively inexpensive, but the term  $m_y$  can also represent 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. In this case we could write  $m_y(a)$  where  $a$  is accuracy and  $m'_y > 0$ , that is, cost is increasing in accuracy.

If the teacher is indifferent as to which lesson to teach, then there is no agency dilemma. The district can 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 best 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 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 will simply present the conclusion.

For sake of argument, let the realization of  $\hat{\rho}_j > \hat{\rho}_k$ , 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

$$\hat{\rho}_j - \hat{\rho}_k > m_y - m_e \quad (2)$$

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 will work well in education. 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, i.e.  $\hat{\rho}_j - \hat{\rho}_k > m_y - m_e$ , while those who oppose output based P4P argue that the reverse is true.

The model shows that the optimal contract depends on the relative costs of monitoring inputs and output and how asymmetrically well-informed the teacher is. The key observation is that if the best lesson is uncertain from the point of view of the district, then payment based on output will be superior to payment based on input. The marginal benefit of delegating the choice of lesson (or the cost of directing the choice of lesson) is the distance of  $\hat{\rho}_j$  from  $\hat{\rho}_k$  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  $\sigma^2$  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 (or cost of choosing the wrong lesson) is large. If there is little technological uncertainty, the expected benefit of the right lesson (or the cost of the wrong lesson) is small.

### 3 P4P in education

The theoretical model 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 K-12 teaching. 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 contracts. I conclude that the status quo is not optimal and consider alternatives.

#### 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 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.<sup>8</sup> 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

---

<sup>8</sup>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.

students opportunities to practice comprehension and reading skills in pairs and small groups.

Figure VI 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. 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 small relative to the dispersion, i.e. there is a lot of overlap in the distributions.<sup>9</sup> Figure VI makes clear that education is characterized by considerable technological uncertainty and thus is the sort of setting where there are potentially large gains from making use of the agent's local knowledge.

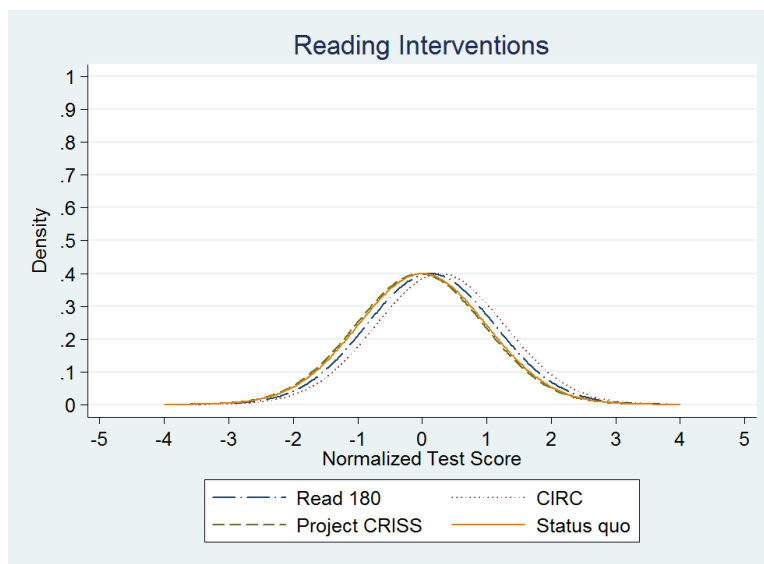


Figure VI: Uncertainty in Education

---

<sup>9</sup>Because I use effect sizes, which force the standard deviation to be 1, I cannot say whether the overlap is due to a large standard deviation or a small difference in means. Recall from the theoretical discussion that what matters is that it is unclear which lesson is best.

### 3.2 Are teachers asymmetrically informed about what is best?

A necessary condition of the model is that the agent be asymmetrically informed. It is not enough for there to be considerable uncertainty, teachers must have local knowledge about which lesson is best. Consider the three interventions in Figure VI. CIRC appears to be the best choice in expectation, but it appears likely that one of the other choices may be better for a particular class.<sup>10</sup> 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 teachers 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  $\hat{\rho}_i$  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 of  $\hat{\rho}_i$  depends 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 figure out what is best for their students - the critical question is whether teachers can figure out 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 make sense of their local knowledge. In the case of 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 describe this in more detail in section 4.

---

<sup>10</sup>A necessary assumption is that class-specific efficacy of the interventions is not correlated. That is, the realizations of  $\hat{\rho}_i$  are independent draws. If this assumption is violated, for example, 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.

### 3.3 What is the nature of monitoring in education?

In section 2 I 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 Education 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”, i.e. incentives for test scores will get you higher test scores but not necessarily real student learning.

Furthermore, even if accurate tests and models were developed, many districts do not have the necessary data infrastructure nor statistical expertise. Additionally, value-added scores are a controversial measure of teacher efficacy that are opposed by powerful interests such as teachers' unions (West and Mykerezzi, 2011). 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 if their attendance or behavior is egregious. Attendance and minimally adequate behavior are measures of input, but they are a low bar to clear. 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.

When pay is contingent on evaluations this is input based P4P.<sup>11</sup> 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” and the likes (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 effectiveness. Measuring inputs on the other hand is already crudely done and there are politically attractive enhancements to this management practice.

### 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

---

<sup>11</sup>However, evaluations are not without drawbacks. 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 based on subjective evaluations done by local staff. Such plans may not improve student achievement because evaluators lack incentives to assess teachers accurately.

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 2007-8 over 65% reported that they have a moderate or great deal of control over “selecting content, topics, and skills to be taught” and over 95% reported that they have a moderate or great deal of control over “selecting teaching techniques”.<sup>12</sup>

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”.<sup>13</sup>

### **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 are usually paid for inputs. I can now ask whether the 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

---

<sup>12</sup>Author’s calculations using the public use version of the 2007-08 SASS.

<sup>13</sup>65% and 95% of teachers in districts that “pay for excellence” 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 perfectly correlated with the district's or the teacher has no preferences other than monetary rewards and it is more costly for the district to measure outputs than inputs.

The model predicts that 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 policy makers. 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 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 uncertainty. Given that most teachers already have control over the curriculum and pedagogy, the model predicts that 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). Specifically, it may raise student scores on the targeted metric. It is not the case, however, that output based P4P in education has been universally successful. Historically, when districts try output based P4P plans, they generally abandon them 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 on the targeted metric 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. If this is the case, schools that implement output based P4P will also need to implement complementary reforms that help teachers make good use of their local knowledge.

### **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 the teacher to “do better” without giving her the resources to figure out what to do. Given one of our most basic tenants in economics - people respond to incentives - is it then any surprise that we see evidence of coaching, gaming, teaching to the test or even cheating?

A teacher who is motivated by P4P will change her behavior to earn the P4P reward. One option is for the teacher to devote time to the aforementioned unproductive hidden actions. Another option is to 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 P4P 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, P4P reforms were 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 find that districts that use this approach to professional development experience gains in student achievement and that districts that couple output based P4P with support for collaboration achieve even greater gains. Consistent with the theoretical model, I find 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.

## 4 Minnesota's Q-Comp

In 2005, the state of Minnesota implemented Quality Compensation (Q-Comp) for teachers 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) sets general guidelines and districts 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.<sup>14</sup>

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

The P4P component is divided into bonuses for teacher 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, \$750 is tied to teacher or small team-level goals and \$250 is tied to school or district-level goals, however there is considerable variation. Histograms and summary statistics are provided in Table I.<sup>15</sup> Payouts for each bonus are generally binary, that is teachers either earn a bonus or they do not. No district has a linear payout scheme and only a few have an option to earn a partial bonus.

Only one district tied rewards to value-added measures. Instead of using value-added

---

<sup>14</sup>The state provides general education aid of approximately \$6,100 per student per year so Q-Comp would add 4% to the average district's baseline funding per student.

<sup>15</sup>See Sojourner et al. (2011) for more detail about the P4P component of Q-Comp.

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.

Teachers earn P4P 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 of both input and output based P4P.

The state encourages districts to link the evaluation P4P 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 2000) which is well regarded in 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 P4P bonuses for participating in the evaluation process and for scoring highly in the observation rubric. This is clearly 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-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-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 a formal level letter 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 the P4P component. Secondly, using a freedom of information act, I obtained additional information on Q-Comp districts. This included data on the state's assessment of how well each district had implemented the agreed upon plan. 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".<sup>16</sup> I construct a measure of the JEPD and evaluation components that measure the percent proficient on the relevant rubric section. More detail on the rubric is provided in the appendix. Histograms and summary statistics of the percent proficient on each component are provided in Table I.

In summary, 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 a decentralized professional development process, JEPD, and the teacher evaluation/observation process.

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. Summary

---

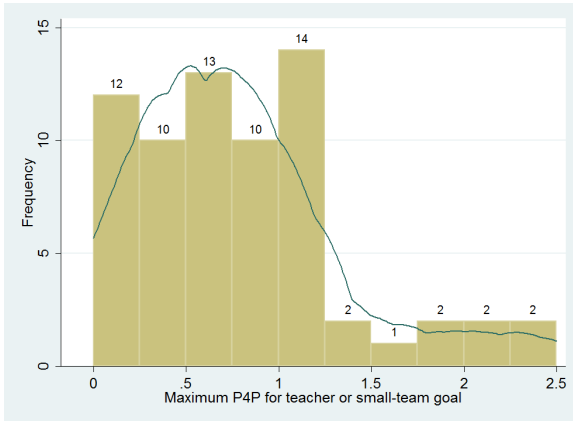
<sup>16</sup>To be exact, the rubric evaluates four of the five components. The alternative salary schedule component is not evaluated. It appears that many districts have struggled with this and that the MDE has deemphasized this component.

statistics are provided in Table II. The resulting panel data provides a unique opportunity to investigate the 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 can inform educational policy as well as provide insights for optimal contracting more generally.

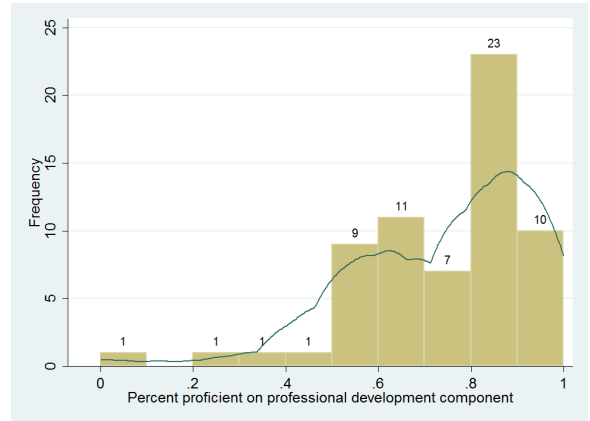
Table I: Descriptive statistics of Q-Comp contracts

Variable	N-districts	Mean	Std. Dev.	Min	Max
Teacher P4P\$	68	.7720294	.5638541	0	2.5
School P4P\$	68	.3783529	.3402682	0	2.5
Evaluation P4P\$	68	.8232353	.5492163	0	2.5
JEPD score	64	.7443173	.2019007	0	1
Evaluation score	64	.6771484	.2133949	.125	1

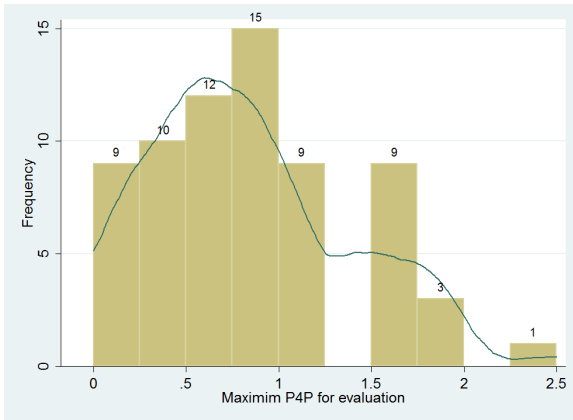
Teacher P4P



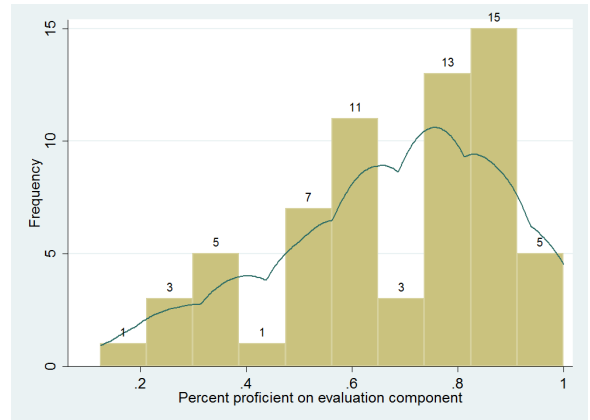
Job-Embedded Professional Development Score



Evaluation P4P



Evaluation Score



School P4P

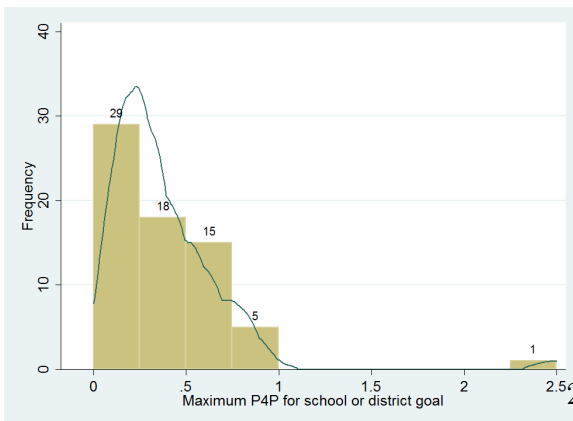


Table II: Descriptive statistics of school and student demographics by year

Variable	Academic year starting				
	2005	2006	2007	2008	2009
Student enrollment	175.699	171.4	169.0	166.6	165.4
	145.222	142.7	139.8	135.4	131.7
Percent free lunch	0.513	0.512	0.511	0.511	0.511
	0.056	0.057	0.058	0.06	0.06
Percent free/reduced	0.228	0.239	0.245	0.253	0.282
	0.184	0.193	0.195	0.197	0.206
Percent special educ.	0.308	0.321	0.327	0.338	0.366
	0.208	0.214	0.216	0.217	0.221
Percent male	0.132	0.134	0.136	0.136	0.138
	0.058	0.059	0.059	0.06	0.062
Percent African-American	0.078	0.087	0.091	0.093	0.095
	0.131	0.14	0.144	0.147	0.146
Percent Hispanic	0.051	0.059	0.062	0.066	0.069
	0.073	0.086	0.091	0.093	0.097
Percent Asian	0.055	0.059	0.061	0.062	0.064
	0.085	0.096	0.1	0.104	0.106
Percent Native American	0.02	0.02	0.02	0.021	0.021
	0.067	0.071	0.071	0.072	0.074

Means and standard deviations listed.

Unit of observation is school level for enrollment and school-grade level for demographics.

Demographics are weighted by count tested (as are regressions).

## 5 Empirical model

I investigate how schools' student achievement changes as their Q-Comp participation changes and test whether schools that pair P4P bonuses with complementary management practices raise student achievement more than schools that do not.

The outcome,  $y_{sgt}$ , is average student achievement on the Minnesota Comprehensive Achievement Series-II tests (MCA-II) for each grade  $g = 3...8$  in school  $s$  in year  $t$ . I report results for math and reading separately.

In explaining school-grade-year average student achievement, I use two specifications:

$$(A) \quad y_{sgt} = \beta_{1t}TeacherP4P_{sgt} + \beta_{1s}SchoolP4P_{sgt} + \beta_{1e}EvalP4P_{sgt} \quad (3)$$

$$+ \beta_2JEPDScore_{sgt} + \beta_3EvalScore_{sgt} \quad (4)$$

$$+ \varphi 1(Dropped_{sgt}) + \lambda 1(Pre - adoption)_{sgt} \quad (5)$$

$$+ \alpha_g w_{sgt} + \gamma_{sg} + \delta_{gt} + \epsilon_{sgt} \quad (6)$$

$$(B) \quad y_{sgt} = \beta_{1t}TeacherP4P_{sgt} + \beta_{1s}SchoolP4P_{sgt} + \beta_{1e}EvalP4P_{sgt} \quad (7)$$

$$+ \beta_2JEPDScore_{sgt} + \beta_3EvalScore_{sgt} \quad (8)$$

$$+ \beta_{12}(TeacherP4P_{sgt} * JEPDScore_{sgt}) \quad (9)$$

$$+ \beta_{13}(EvalP4P_{sgt} * EvalScore_{sgt}) \quad (10)$$

$$+ \varphi 1(Dropped_{sgt}) + \lambda 1(Pre - adoption)_{sgt} \quad (11)$$

$$+ \alpha_g w_{sgt} + \gamma_{sg} + \delta_{gt} + \epsilon_{sgt} \quad (12)$$

I describe a Q-Comp district's P4P using a vector of three variables that measure the maximum bonus available for teacher or small-team level goals,  $TeacherP4P$ , school or district level goals,  $SchoolP4P$  and formal evaluations,  $EvalP4P$ . I describe a Q-Comp's management components using the percent proficient on the state's rubric for the job-

embedded professional development component, *JEPD*, and the evaluation component, *EvalScore*. In specification (B), I test the complementarity of these management practices with *TeacherP4P* and *EvalP4P* respectively.<sup>17</sup> There is no corresponding management reform that supports *SchoolP4P*.

Each of these is indexed by school-grade-year, however, in most cases the district's Q-Comp program is the same across all school-grades in participating years. The few exceptions were coded appropriately.

If  $\beta_{1t} > 0$ ,  $\beta_{1s} > 0$ , and/or  $\beta_{1e} > 0$  then teacher-level, school-level or evaluation based P4P leads to increased student achievement when implemented alone. If  $\beta_2 > 0$  and/or  $\beta_3 > 0$ , job-embedded professional development and/or formal evaluations lead to increased student achievement when implemented alone.

Specification (B) includes two important interactions. First,  $TeacherP4P_{sgt} * JEPD_{sgt}$  measures the joint impact of teacher or small-team level P4P and job embedded professional development. A positive coefficient would indicate that the two are complements and best implemented together. More exactly, if the effect of rewarding teacher level goals is greater when the JEPD is stronger (or the effect of JEPD on student outcomes is greater when more dollars are at risk for teacher level goals) then  $\beta_{12}$  will be positive. Second,  $EvalP4P_{sgt} * Evalscore_{sgt}$  measures the joint impact of linking P4P to evaluations and the strength of the district's management policies regarding evaluation. A positive coefficient would indicate that the two are complements and best implemented together. More exactly, if the effect of rewarding a good score on an evaluation 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  $\beta_{13}$  will be

---

<sup>17</sup>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 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.

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 estimates of  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$ ,  $\beta_{12}$  and  $\beta_{13}$ . To guard against this, I include school-grade-year student demographics  $w_{sgt}$  to control for time-varying observable differences, and school-grade fixed effects ( $1_{sg}$ ) to control for time-invariant unobservable differences. The model is identified from within-school-grade, across-time variation. Fixed effects for each year-grade ( $1_{gt}$ ) are also included. These terms identify counter-factual year effects for each grade. 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.<sup>18</sup>

I also include an indicator for school-grades that were in Q-Comp but have since dropped out. I include an indicator for academic years two or more years prior to adoption, which I label “pre-adoption”. This conditions on and measures pre-adoption differences in achievement levels between Q-Comp districts and non-Q-Comp districts. The reference category is the single year immediately prior to adoption. A negative coefficient would indicate 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 2011).

---

<sup>18</sup>The first difference is the within-school comparison across time periods. The second difference is between the first-differences at adopting schools and those at non-adopting schools across the same time period. A within-school change between any two points in time is evaluated against changes across those same two years among other schools.

## 6 Results

### 6.1 Reading

The results that use school-grade average scores on the MCA-II reading exam as the outcome are reported in Table III. In both specifications (A) and (B) rewards for teacher or small team-level goals results in an increase in student achievement on the order of  $0.1\sigma$ . This result is significant at the 1% level. No other P4P category is associated with student achievement gains (or losses). In specification (B) this should be interpreted as the conditional relationship between rewarding teacher-level goals and JEPD. In this case, the coefficient on *TeacherP4P*\$ is the effect of a \$1,000 bonus when a district does not score proficient on *any* of the JEPD elements of the rubric (i.e. *JEPDScore* = 0).

The results also show that higher score on the JEPD component of the rubric is associated with gains in student achievement. This result is significant at the 10% level in specification (A) and remains positive, although it is smaller in magnitude and is no longer statistically significant at conventional levels in specification (B). Again, specification (B) should be interpreted with care. The coefficient on *JEPDScore* is 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).

The main coefficient of interest, the interaction term between *TeacherP4P*\$ and *JEPDScore*, is positive but not significant. There is weak evidence that the impact of rewarding teacher-level goals is increasing in the strength of the complementary JEPD.

For reading, Q-Comp appears to have had its largest impact via the P4P bonuses attached to teacher-level goals. Districts that implemented strong JEPD may have experienced gains as well and the two may be complements, however, the statistical evidence is not strong.

Interestingly, districts that score higher on the evaluation component of the rubric experience *declines* in reading scores. This result is significant at the 10% level in specification (A) and retains its negative sign in specification (B), although it is smaller in magnitude and

no longer statistically significant at conventional levels. The interaction of  $EvaluationP4P\$$  and  $EvalScore$  is negative but not significant. It does not appear that linking bonuses to evaluations nor the practice of conducting formal evaluations leads to gains in reading achievement either individually or in tandem.

The coefficient on the pre-adoption indicator is not significant, however, the indicator for districts that were once in Q-Comp but have since dropped out is significant and negative. This indicates that these districts experienced declines in reading achievement after dropping Q-Comp. All observable student demographics such as race and gender are significant and of the expected sign and are not reported here for brevity but are available upon request.

## 6.2 Math

For math, the effect of P4P for teacher-level goals is *negative*, although it is small in magnitude and statistically insignificant in specification (A). The coefficient -0.037 in specification (B) indicates that districts which reward teacher-level goals but do not score proficient on any part of the JEPD component of the rubric, experience *declines* on the order of  $0.037\sigma$  on standardized math assessments.

On the other hand, districts that scored highly on JEPD component of the rubric experienced large gains in math achievement. In specification (B) we see that districts that score highly on the JEPD component of the rubric, but do not attach rewards to teacher-level goals, experience gains on the order of  $0.318\sigma$ . The interaction term is also positive and statistically significant. The impact of JEPD appears to be even greater when more money is at risk for teacher-level goals. This is shown by the positive and significant 0.094 result on  $TeacherP4P\$ * JEPDScore$ .

For math, it seems that Q-Comp has its largest impact via the management practices that support collaboration. The districts that were most successful also tied bonuses to the goals that were set via this decentralized professional development process thus there is evidence that the two are complements.

Table III: Effects of Q-Comp on Reading Achievement

	MCAII Reading - Average Score	
	(A)	(B)
Teacher P4P\$	0.119*** (0.029)	0.106*** (0.018)
School P4P\$	-.039 (0.081)	-.086 (0.088)
Evaluation P4P\$	-.048 (0.03)	0.015 (0.097)
JEPD score	0.228* (0.118)	0.142 (0.143)
Evaluation score	-.264* (0.137)	-.163 (0.169)
Teacher P4P\$ * JEPD score		0.038 (0.054)
Eval P4P\$ * Eval score		-.090 (0.133)
1(Pre-adoption)	-.026 (0.049)	-.029 (0.05)
1(Dropped)	-.103*** (0.027)	-.106*** (0.027)
Student observables	Yes	Yes
District-grade FE	Yes	Yes
Year-grade FE	Yes	Yes
<i>N</i> districts	471	471
<i>N</i> school-grades	4676	4676
<i>N</i> tested students	1749818	1749818
Adj. R <sup>2</sup>	0.887	0.887

Standard errors calculated using a robust cluster variance estimator.

*N* school-grades represents the number of clusters.

*N* tested students represents the number of school-grade-years weighted by count tested.

The effect of rewarding evaluations is again *negative*. Districts that scored highly on the evaluation component of the rubric experienced large *declines* in math achievement. However, unlike the results from the reading outcomes, the impact of evaluations appears to reverse when more money is at stake based on the result of the evaluation. The positive and significant 0.215 on  $EvalP4P\$ * EvalScore$  indicates that districts that have larger rewards for evaluations do not experience as strong a decline as districts that focus on evaluations without corresponding P4P rewards.

For math, as in reading, the pre-adoption indicator is not significant but districts that drop Q-Comp appear to experience significant declines. Observable student demographics are again excluded from the table but are of expected sign and are available upon request.

For readers unfamiliar with the literature on educational outcomes, it is worth noting that the magnitudes of the estimates are quite large. For comparison, Krueger (1999) estimates that Project STAR, which reduced class sizes in Tennessee, resulted in gains on the order of  $0.2\sigma$ . The social value of a  $0.2\sigma$  gain for teacher's class each year has been recently estimated conservatively at \$200,000 (Hanushek, 2010; Chetty et al., 2010).

Table IV: Effects of Q-Comp on Math Achievement

	MCAII Math - Average Score	
	(A)	(B)
Teacher P4P\$	-.018 (0.025)	-.037** (0.017)
School P4P\$	0.016 (0.114)	0.07 (0.103)
Evaluation P4P\$	-.040 (0.032)	-.155*** (0.047)
JEPD score	0.268*** (0.094)	0.318*** (0.087)
Evaluation score	-.221** (0.094)	-.410*** (0.111)
Teacher P4P\$ * JEPD score		0.094* (0.057)
Eval P4P\$ * Eval score		0.215*** (0.079)
1(Pre-adoption)	0.005 (0.131)	0.007 (0.132)
1(Dropped)	-.082** (0.034)	-.079** (0.032)
Student observables	Yes	Yes
District-grade FE	Yes	Yes
Year-grade FE	Yes	Yes
<i>N</i> districts	469	469
<i>N</i> school-grades	4665	4665
<i>N</i> tested students	1698331	1698331
Adj. R <sup>2</sup>	0.86	0.86

Standard errors calculated using a robust cluster variance estimator.

*N* school-grades represents the number of clusters.

*N* tested students represents the number of school-grade-years weighted by count tested.

## 7 Discussion

In this paper I present a principal-agent model adapted from Prendergast (2002). I provide evidence that education fits the assumptions of the model and offer an empirical test using evidence from Minnesota’s Q-Comp program. I find both theoretical and empirical support for the complementarity of P4P and decentralized decision making.

I present a model that is 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 causally dealt with. Importantly, I identify the assumptions needed for output based P4P to work in education and suggest complementary reforms.

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 that best fits a particular set of students and a particular teacher is selected.

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 complementary reforms, specifically time and support for collaboration, to help teachers make sense of their local knowledge.

Opponents of output based P4P who are concerned about dangerous hidden actions will see these concerns accounted for in the model. Opponents who argue that output based P4P is demeaning to teachers should consider 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, i.e. time and opportunities for collaboration, that support teachers in making good use of their local knowledge.

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.

Of course, if teachers have preferences that are perfectly correlated with the district, it may be enough to offer more opportunities for collaboration. I offer empirical evidence to the contrary. I find that in P4P bonuses for reading have a larger impact than support for collaboration and in math, support for collaboration is strengthened by P4P. I conclude that the two are complements and best implemented together. 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.

## References

- P. Aghion and J. Tirole. Formal and real authority in organizations. *Journal of Political Economy*, 105(1):1–29, 1997.
- A. Atkinson, S. Burgess, B. Croxson, P. Gregg, C. Propper, H. Slater, and D. Wilson. 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, 2004.
- G. Baker, R. Gibbons, and K.J. Murphy. Subjective performance measures in optimal incentive contracts. *The Quarterly Journal of Economics*, 109(4):1125–1156, 1994.
- N. Bloom and J. Van Reenen. Human resource managment. *Handbook of Labor Economics*, 4b:1697–1763, 2011.
- R. Chetty, J.N. Friedman, N. Hilger, E. Saez, D.W. Schanzenbach, and D. Yagan. How does your kindergarten classroom affect your earnings? evidence from project star. Technical report, National Bureau of Economic Research Working Paper No. 16381, 2010.
- D.N. Figlio and J. Winicki. Food for thought: the effects of school accountability plans on school nutrition. *Journal of Public Economics*, 89(2-3):381–394, 2005.
- R.G. Fryer. Teacher incentives and student achievement: Evidence from new york city public schools. Technical report, National Bureau of Economic Research Working Paper No. 16850, 2011.
- S. Grossman and O. Hart. The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy*, 94(4):691–719, 1986.
- E.A. Hanushek. The economic value of higher teacher quality. *Economics of Education Review*, 2010.

- E.A. Hanushek and S.G. Rivkin. Teacher quality. *Handbook of the Economics of Education*, 2:1051–1078, 2006.
- B. Holmstrom and P. Milgrom. Multitask principal–agent analyses: incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and organization*, 7(special issue):24, 1991.
- B.A. Jacob. Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics*, 89(5-6):761–796, 2005.
- B.A. Jacob and L. Lefgren. Can principals identify effective teachers? evidence on subjective performance evaluation in education. *Journal of Labor Economics*, 26(1):101, 2008.
- B.A. Jacob and S.D. Levitt. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, 118(3):843, 2003.
- A. Krueger. Experimental estimates of education production functions. *The Quarterly Journal of Economics*, 114(2):497–532, 1999.
- J. R. Lockwood, D. McCaffrey, L.S. Hamilton, B. Stecher, V. Le, and J.F. Martinez. The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, 2007.
- P.S. Martins. *Individual teacher incentives, student achievement and grade inflation*. IZA Discussion Paper No. 4051, 2009.
- R. Murnane and D. Cohen. Merit pay and the evaluation problem: Why most merit pay plans fail and few survive. *Harvard Educational Review*, pages 1–18, 1986.
- D. Neal. The design of performance pay in education. *Handbook of the Economics of Education*, 4:495–550, 2011.
- C. Prendergast. The tenuous trade-off between risk and incentives. *Journal of Political Economy*, 110(5), 2002.

- R. Radner. The organization of decentralized information processing. *Econometrica*, 61(5): 1109–1146, 1993.
- J.E. Rockoff, B.A. Jacob, T.J. Kane, and D.O. Staiger. Can you recognize an effective teacher when you recruit one? *Education Finance and Policy*, 6(1):43–74, 2011.
- J. Rothstein. Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, 25(1), 2010.
- A. Sojourner, K.L. West, and M. Mykerezzi. When does teacher performance pay raise student achievement: Evidence from minnesota’s q-comp program. 2011.
- M.G. Springer, L. Hamilton, D.F. McCaffrey, D. Ballou, V.N. Le, M. Pepper, JR Lockwood, and B.M. Stecher. Teacher pay for performance: Experimental evidence from the project on incentives in teaching. *National Center on Performance Incentives. Peabody College of Vanderbilt University, PMB No. 43, 230 Appleton Place, Nashville, TN 37203. Tel: 615-322-5538; Fax: 615-322-6018; e-mail: ncpi@vanderbilt.edu; Web site: http://www.performanceincentives.org*, 2010.
- J.H. Tyler, E.S. Taylor, T.J. Kane, and A.L. Wooten. Using student performance data to identify effective classroom practices. *American Economic Review*, 100(2):256–60, 2010.
- K.L. West and E. Mykerezzi. Teachers’ unions and compensation: The impact of collective bargaining on salary schedules and performance pay schemes. *Economics of Education Review*, 30(1):99–108, 2011.

## 8 Appendix

### 8.1 Model Details

The timing is of the model such that the district and the teacher first agree to a contract. The teacher learns  $\rho_i$  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. We solve the problem using backward induction.

The district seeks to maximize student achievement net costs. Equation (13) describes the district’s problem.

$$\text{Max}(E[y_i - w_i - m]) \tag{13}$$

The maximization is in expectation because the district does not know  $\rho_i$  *ex ante* when it agrees to a contract (recall that  $y_i$  is a function of  $\rho_i$ ).

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 (13) 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 (14) describes the teacher’s problem.

$$\text{Max}(w_i - C(e_i) + B) \tag{14}$$

For simplicity, assume that the teacher is risk neutral.<sup>19</sup> 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

---

<sup>19</sup>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.

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 and  $C(e_j) > C(e_i)$  for some lesson  $j$ . In equation (14) I let one lesson cost  $C(e) - B$  and every other lesson cost  $C(e)$ .

### 8.1.1 Contract type: Input - Delegate

If the district delegates the choice of lesson to the teacher and pays based on inputs, the expected output is  $E[y_i] = (\sum_{i=1}^n (\bar{\rho}_i + e_i^*)) / n$ , where  $e_i^*$  is the optimal level of effort obtained from the teacher's maximization problem. 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 cost of effort reduced by  $B$  so  $w_i = \sum_{i=1}^n C(e_i^*) - B$ .

The district also incurs monitoring costs,  $m_e$ , so the resulting surplus is described by equation (15).

$$E[y_i - w_i - m] = (\sum_{i=1}^n (\bar{\rho}_i + e_i^*) / n) - (\sum_{i=1}^n C(e_i^*) - B) - m_e \quad (15)$$

### 8.1.2 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 (16).

$$E[y_i - w_i - m] = (\bar{\rho}_k + e_k^*) - (C(e_k^*) - B/n) - m_e \quad (16)$$

Equation (16) > equation (15) 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 about the same, on average, in which case it can let the teacher choose the lesson and pay a slightly lower wage to capture some of teacher's rents.

### 8.1.3 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 (17).

$$E[y_i - w_i - m] = \bar{\rho}_k + e_k^* - (C(e_k^*) - B/n) - m_y \quad (17)$$

Equation (16) > equation (17) because of the assumption that  $m_y > m_e$ .

### 8.1.4 Contract type: Output - Delegate

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.

For sake of argument, let the realization of  $\hat{\rho}_j > \hat{\rho}_k$ , 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 (18) describes the expected surplus.

$$E[y_i - w_i - m] = \rho_j + e_j^* - (C(e_j^*) - B/n) - m_y \quad (18)$$

To see if equation (18) > equation (16), we need to know how much greater  $\hat{\rho}_j$  is than

$\hat{\rho}_k$ . 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 (18) > equation (16) 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  $\sum_{i=1}^n \bar{\rho}_i + e_i^*/n = (\sum_{i=1}^2 \bar{\rho}_i + e_i^*)/2 = 0 + e_i^*$ . 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 statistics of two draws from a normal distribution with mean 0 and variance  $\sigma^2$ . Therefore rewrite equation (18) as:

$$E[y_i - w_i - m] = \sigma/\sqrt{\pi} + e_j^* - C(e_j^*) + B/n - m_y \quad (19)$$

noting that  $\bar{\rho}_k = 0$  in this example, (19) > (16) reduces to:

$$\sigma/\sqrt{\pi} > m_y - m_e \quad (20)$$

The inequality (20) 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 (20) holds. This is likely if the left hand side is big (lots of uncertainty) and/or the right hand side is small. Since we have assumed that  $m_y$  includes the “costs” of multitasking, what equation (20) 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. This conclusion is discussed at more length in the main text.

## 8.2 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 sub-sections. 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 (*Evaluation score* = 0.57). Districts that have a higher percent proficient have implemented the evaluation process with more fidelity.