Title
Essays on Personnel Economics in Low-Income Countries

Author
Brown, Christina L

Publication Date
2021

Peer reviewed|Thesis/dissertation

Essays on Personnel Economics in Low-Income Countries

by

Christina L. Brown

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Supreet Kaur, Chair
Professor Edward Miguel
Professor Christopher Walters

Spring 2021

Essays on Personnel Economics in Low-Income Countries

Abstract

Essays on Personnel Economics in Low-Income Countries

by

Christina L. Brown

Doctor of Philosophy in Economics

University of California, Berkeley

Professor Supreet Kaur, Chair

A key question in personnel economics is how best to motivate and incentivize workers. In this dissertation, I investigate how different incentive systems affect workers' effort and decision on where to work. Rewarding different aspects of workers' performance may allow firms to prioritize certain outcomes and may attract and retain different types of employees who are more or less drawn to particular contracts. Finally, certain incentive schemes may benefit or harm certain sub-groups of employees, especially when there is subjectivity introduced into the evaluation scheme.

In the first chapter, joint with Tahir Andrabi, we study whether performance incentives lead to sorting of teachers. Attracting and retaining high-quality teachers has a large social benefit, but it is challenging for schools to identify good teachers ex-ante. We use teachers' contract choices and a randomized controlled trial of performance pay with 7,000 teachers in 243 private schools in Pakistan to study whether performance pay affects the composition of teachers. Consistent with adverse selection models, we find that performance pay induces positive sorting: both among teachers with higher latent ability and among those with a more elastic effort response to incentives. Teachers also have better information about these dimensions of type than their principals. Using two additional treatments, we show effects are more pronounced among teachers with better information about their quality and teachers with lower switching costs. Accounting for these sorting effects, the total effect of performance pay on test scores is twice as large as the direct effect on the existing stock of teachers, suggesting that analyses that ignore sorting effects may substantially understate the effects of performance pay.

In the second chapter, joint with Tahir Andrabi, we investigate how different types of incentive pay affect employee behavior. A central challenge facing schools is how to incentivize teachers. While high-powered incentives can motivate effort, they can lead teachers to distort effort away from non-incentivized outcomes. This is one reason why most performance incentives allow for manager subjectivity. However, this subjectivity can introduce new concerns, including favoritism and bias. We study the effect of subjective

versus objective performance incentives on teacher productivity using the same randomized controlled trial discussed in chapter 1. We estimate the effect of two performance raise treatments versus a control condition, in which all teachers receive the same raise. The first treatment arm is a "subjective" raise, in which principals evaluate teachers; the second treatment arm an "objective" raise based on student test scores. First, we show that both subjective and objective incentives are equally effective at increasing test scores. However, objective incentives decrease student socio-emotional development. Second, we show that these effects are likely driven by the types of behavior change we observe from teachers during classroom observations. In objective schools, teachers spend more time on test preparation and use more punitive discipline, whereas, in subjective schools, pedagogy improves. Finally, we investigate the mechanisms of these effects through the lens of a moral hazard model with multi-tasking. We exploit variation within each treatment to isolate the causal effect of contract noisiness and distortion on student outcomes. We then show that teachers perceive subjective incentives as less noisy and less distorted, and these contract features affect student outcomes, serving as key channels to explain the reduced form effects we see.

Finally, in the third chapter, I explore whether managers show gender bias in their evaluation of employees, and, if so, under what circumstances. Pakistan ranks in the lowest decile in female labor force participation, and even in sectors where women are more prevalent, such as teaching, they earn 70 cents for each dollar men earn. In this chapter, I test the extent to which statistical versus financial discrimination explains these pay gaps. I use the experiment from chapter 1 and 2, which has two important random variations: i). how often managers observe a given employee and ii). whether manager evaluations affect employee's pay or are just used for feedback and see whether this changes how managers evaluate their employees. I find that managers have less gender bias the more frequently they observe a given employee and more gender bias if there is a financial stake of the manager's evaluation.

While all three chapters use the same randomization design and data, each chapter is intended to be a stand-alone set of research questions, so the respective design and data description is included within each chapter.

To my family –
my parents, Louise and Paul,
my sister, Marissa,
and my husband, Michael –
for believing in me when I didn't.

Contents

List of Figures

List of Tables

Acknowledgments

There are so many people that have given their time, energy, and resources to support me and this project over the years. I could write a whole dissertation just about their generosity.

First, I am so thankful to my committee, Supreet Kaur, Ted Miguel, and Christopher Walters. Supreet reshaped how I think about labor economics. She exposed me to part of the labor market I didn't know existed and has done so much to expand our understanding of low-income workers, especially in South Asia. I have also learned so much about how to think about economic insights and turn that into a practical, well-designed experiment from her. She has always pushed me to grow and never accepted less than my best.

Ted has had a huge influence on my career that extends years before we ever met. I read a number of his papers prior to starting my PhD. The example he set in conducting policy-relevant development research was one of the reasons I chose to pursue an economics PhD. Since coming to Berkeley, Ted has been a source of endless encouragement and advice.

Chris is equal parts brilliant and kind, which is unusual for characteristics that are generally negatively correlated in this profession. I often felt like an imposter in the world of labor economics but with Chris that was never an issue. He has been patient, attentive, and caring, and his research is an exemplar of the power economics can have in improving public policy.

In addition to my committee, many other individuals have provided guidance and support. Asim Khwaja has influenced my work in more ways than is possible to enumerate. He is the person I have learned the most from and is an endless source of insight, humor, and energy. I aspire to be 1% of the economist he is. Tahir Andrabi has been one of my biggest advocates for a long time, and this dissertation simply would not have existed without his support, generosity, and kindness. I feel extremely lucky to have him as a co-author. Fred Finan was a source of honest, terribly insightful advice and one of my favorite people to bounce ideas off. I am a better economist because of him. I also would like to thank David Card, Jishnu Das, Stefano DellaVigna, Patrick Kline, Jeremy Magruder, Gautam Rao, Jesse Rothstein, and Heather Schofield for providing advice and encouragement over the last six years.

I am eternally indebted to Sam Leone, Peter McCrory, and Preston Mui for helping me survive, and even enjoy, first year and beyond. Your friendship has been a buoy throughout the ups and downs of graduate school. I am so lucky to have you in my life. I have learned so much and immensely benefited from my friends and colleagues, Chris Campos, Luisa Cefala, Isabelle Cohen, Ingrid Haegele, Junyi Hou, Eric Hsu, Anne Karing, Julien Lafortune, Todd Messer, Max Mueller, Jonathan Schellenberg, and Avner Strulov-Shlain. Finally, Patrick Allen has been a ray of sunshine throughout my six years at Berkeley. From day one, he made me feel so welcome.

Funding for this project was generously provided by DFID's RISE Programme, JPAL's Post-Primary Initiative, the Weiss Family Fund, CEGA, the Strandberg Fund, the National Academy of Education/Spencer Foundation, and the Institute for Research on Labor and Employment.

I have had the extreme pleasure of working with the Center for Economic Research in Pakistan for the last seven years. They make flying halfway across the world feel like coming to my second home. I cannot thank Haya Mubasher, Anam Tariq, Attefaq Ahmed, Zahra Niazi, Mujahid Murtaza, Maheen Rashid, and Zohaib Hassan enough for their incredible research assistance and Wasif Mullick and Faisal Riaz for their help throughout this project.

I also owe a huge debt to the anonymous school system, which I partnered with for this experiment. I have learned so much from conversations with these dedicated professionals and am so grateful they trusted me to collect this data.

Lastly, I owe an endless thank you to my family – my parents, Louise and Paul, my sister, Marissa, and my husband, Michael. Through ups and downs, they've supported me unconditionally, cheered for me during the ups and comforted me during the downs. My love of math and social issues (which is essentially what economics is) started from a very young age, and I know that is from the values my parents instilled in me. Marissa has been a source of joy, laughter, and love for my whole life. Finally, none of this would be possible without Michael. From solving my latex errors to being a shoulder to cry on to moving across the country twice, his love and humor make every day better.

Chapter 1

Inducing Positive Sorting Through Performance Pay

1.1 Introduction

Teachers are the most important input in the education production function, but schools imperfectly observe teacher quality, making it hard to effectively screen teachers. The characteristics available to schools, such as experience, college grades, credentials, and interview scores, are poor predictors of future performance, explaining less than 5% of the variation in teacher value-added [Bau and Das, 2020, Staiger and Rockoff, 2010]. This challenge is not unique to schools. The majority of firms cite challenges in hiring and retaining high-quality employees [The World Bank Group, 2019].

Incentive contracts offer a potential solution to this problem. Even if employers cannot identify employee quality directly, high performers will sort into firms that offer performance pay if employees have private information about their ability. Performance incentives have become increasingly common in teaching, and currently, two-thirds of countries offer performance incentives to public school teachers [The World Bank Group, 2018]. While we have a substantial body of evidence on the effect of performance pay for the existing stock of teachers, we know much less about whether performance pay could induce positive sorting.

In this paper, we use a large-scale experiment to answer three questions: Does performance pay induce positive sorting among teachers? How much asymmetric information is there between schools and teachers? What affects the magnitude of positive sorting? Our experiment is informed by a Roy-style model of job choice in which employers offer different contracts, and employees choose where to work based on their information about their type. We partner with a network of private schools located in urban Pakistan, randomly assigning performance pay among 243 schools.

Our experiment proceeds in two phases. First, we offer teachers the opportunity to choose their contract for the coming year, selecting between a flat raise versus a performance-based raise. Teachers' choices are implemented in a randomly selected subset of schools to ensure

incentive compatibility of responses. We also elicit the distribution of teachers' beliefs about their value-added and risk preferences through an incentivized activity.

Second, among the remaining schools that were not assigned to implement the teacher's choice, we randomize contracts across schools. Teachers receive a flat raise (guaranteed irrespective of performance) or a performance raise (based on student test score performance or principal evaluation). Teachers are informed that the contract type is associated with the school itself, which is important in this setting, as 15% of teachers transfer to work at a different school each year. We then observe what types of teachers move into schools assigned flat versus performance raise contracts over the next year.

We draw on administrative data, baseline and endline surveys of teachers and principals, endline student tests and surveys, and detailed classroom observation data from 7,000 teachers and 50,000 students. Combined, these data allow us to measure teacher value-added and effort along numerous dimensions. We also capture teachers' beliefs about their quality and principal evaluations of teachers along various metrics. Finally, we measure several dimensions of teacher preferences and characteristics, including risk, pro-sociality, and career ambition.

Overall, we find strong evidence that performance pay induces positive sorting among high performing teachers. First, we find that teachers who choose performance pay contracts have higher value-added. Contract choice is predictive of value-added even when controlling for principal's information about teachers. These results are strongest among teachers in the middle of their careers (6-10 years of experience). Second, we find positive sorting along actual job choice. The composition of teachers in schools assigned to performance pay is better after one year. These effects are mostly driven by high value-added teachers moving from control to treatment schools and low value-added teachers moving from treatment to control schools. High value-added teachers are also slightly more likely to leave control schools to work outside this network of schools. We do not find any effect on new entrants to the school system.

Teachers also positively sort on their behavioral response to incentives. Teachers who chose performance pay contracts during the baseline choice exercise have nearly nine times the effect of performance pay on test scores as those who chose flat pay. Moreover, the treatment effect is not correlated with baseline value-added, suggesting that these two aspects of teacher type are unrelated. If we take into account the sorting effects on both value-added and behavioral response, the total effect of performance pay on test scores is nearly twice as large as when we just measure the behavioral effects on the existing stock of employees.

While it is useful to understand whether teachers have information about their type along these two dimensions, part of the sorting value of the incentive contracts depends on whether teachers have *private* information about their type beyond what their employer knows. We find that all our key results hold when we control for principals' evaluations of teachers. Principals do have some information about teacher quality, and they are especially good at rating teachers along highly observable criteria like attendance and behavioral management of students. However, teacher's contract decisions are three times as predictive of value-added as information available to schools (credentials, experience, age, and principal evaluation). This asymmetric information between teachers and principals holds for all except very novice

teachers.

We use two additional sources of random variation to show that the extent of positive sorting varies substantially by teachers' information and switching costs. We randomize teachers to receive information about their value-added from the previous year during the contract choice exercise. This results in a significant improvement in teachers' priors of their future value-added, and a stronger relationship between teacher's value-added and whether they chose a performance pay contract. We also compare teacher's sorting across schools for teachers who have higher versus lower switching costs. We exploit exogenous variation in switching costs by comparing teachers whose closest neighboring school received the opposite treatment status (low switching cost) versus the same treatment status (high switching cost) as their own school. There is four times more positive sorting under low switching costs. This suggests that the extent of positive sorting depends on the ease at which teachers can change jobs in response to incentive contracts.

Our last reduced form result shows that performance pay does not generate sorting of "bad" types into performance pay schools. Surprisingly, teachers who chose performance pay are much less likely to exhibit distortionary behaviors in response to performance incentives than those who chose flat pay. Performance pay also increases other areas of student socio-emotional development for teachers who chose the contract. This suggests that teachers who sort in are not solely focused on maximizing their salary at the cost of more well-rounded student development. Lastly, we do not find evidence that teachers who chose performance pay have other negative traits. They are slightly more likely to contribute to school public goods, to collaborate with other teachers, and have similar levels of pro-sociality (measured using a volunteer opportunity task).

Finally, we use the estimates of teacher's priors, distribution of ability and behavioral response, and elasticity of supply to a given job from our experiment to estimate the effects of a longer-term performance pay policy applied to a larger set of schools. We find that introducing a 30-year performance pay policy (20% of teacher's base salary) across all schools would result in effects of 0.09 SD - 0.17 SD each year. These effects are 1.3-2.4x larger than the one-year effect of performance pay, which only includes the behavioral effect.

Our paper makes three key contributions to the literature. It is the first study to show that performance pay contracts induce positive sorting among existing teachers. We build on a growing literature on understanding the effect of different contract types on teacher selection, the closest of which are two studies that show higher value-added teachers choose performance pay when they are given the option in a low-stakes and high-stakes settings [Johnston, 2020, Leaver et al., 2019]. Related work by Biasi [2017] and Rothstein [2015] provide empirical and structural evidence for the effect of different types of contracts on teacher sorting. There is also an extensive theoretical and empirical literature on adverse selection and performance pay in other sectors [Lazear, 2000, Akerlof, 1970, Lazear and Moore, 1984].

Second, we add to a robust literature on the direct, behavioral effect of performance pay for teachers by providing two new findings [Lavy, 2009, Muralidharan and Sundararaman, 2011, Fryer, 2013, Goodman and Turner, 2013]. We show that there is substantial heterogeneity in the direct effect of performance pay across teachers. Specifically, teachers

who want performance pay have much larger behavioral responses than those that do not want performance pay. This suggests that in the long run, the effects of performance incentives could be much larger than the short term effects previously estimated. In addition, this behavioral response appears to be unrelated to baseline value-added. This suggests that the marginal effort response to incentives is uncorrelated with the equilibrium effort under no incentives.

Third, we isolate the factors which influence the extent of positive sorting. We show the first evidence that higher switching costs dampen the extent of positive sorting, and employee private information increases positive sorting. These results are in line with a rich body of theoretical work on adverse selection [Akerlof, 1970, Lazear and Moore, 1984, Greenwald, 1986] and help us understand the variation in sorting effect sizes across several existing empirical papers [Lazear, 2000, Leaver et al., 2019, Biasi, 2017].

The remaining sections are organized as follows: Section 1.2 provides context about the use of performance pay in teaching. Section 1.3 presents the motivating model in the vein of Roy [1951]. Section 3.3 details the contract choice elicitation, randomized controlled trial, and data collection procedures. Section 1.5 presents the results on the extent of positive sorting in response to performance pay, and Section 1.6 describes the extent of information principals have about teachers. Section 1.7 presents results on the sensitivity of the magnitude of positive sorting to teacher's switching costs and information, and Section 1.8 examines whether there is sorting along negative characteristics. Section 1.9 presents results from a policy simulation exercise.

1.2 Teacher Quality, Labor Market and Performance Pay

Many students in developing countries experience sub-par teaching. In Pakistan, teachers are only present 89% of the time, and 20% of children cannot read a sentence in the local language or solve a two-digit subtraction problem by the end of fifth grade [ASER, 2019]. These patterns are consistent across many low-income countries [Group, 2018]. The dearth of good teaching has large, long-lasting, and diverse negative consequences for students. In Pakistan, exposure to a 1 standard deviation (SD) better teacher results in 0.15 SD higher test scores [Bau and Das, 2020]. There is substantial evidence on the long-term benefits of teacher quality in the US, on a wide array of outcomes from income to crime [Chetty et al., 2014b, Jackson, 2018, Rose et al., 2019].

Despite the importance of teacher quality, schools have limited capacity to screen in and retain good teachers and screen out and lay-off bad teachers, due to institutional and information constraints. Public schools are typically severely constrained in their ability to fire bad teachers. Furthermore, it is not clear that schools can even identify who the high and low performing teachers are, either at the time of hiring or throughout the teacher's tenure. Characteristics available to schools at the time of hiring, including interview scores, explain less than 5% of teacher value-added [Bau and Das, 2020, Staiger and Rockoff, 2010,

Rockoff and Speroni, 2010]. Schools could potentially exploit teachers' private information about their quality by offering performance pay and causing high-quality teachers to self-select in. Lazear [2000] shows that employees in a glass factory positively sort in response to performance pay, and sorting effects are twice as large as the effects on effort.

It is unclear whether we would see more or less asymmetric information in teaching, relative to manufacturing. It is likely harder for employers to assess productivity in higher-skilled professions, like teaching, which have a complicated production function. However, teacher performance pay is generally constructed using an opaque performance incentive metric (typically value-added), and teachers may have little information about their own performance along this metric. Springer et al. [2010] find no relationship between teachers' prediction of whether they will receive a performance-based bonus and actual teacher performance. At baseline, we also ask teachers to predict their rank along the performance metric. We also find no relationship between teachers' predictions and actual performance. However, these low-stakes survey questions may not reflect the true extent of information teachers have.

Understanding the full effects of performance pay including both direct effects on existing teachers and sorting effects is crucial, as there has been a significant push to tie teacher salaries to student outcomes in developed and developing countries [Goodman and Turner, 2013, Pham et al., 2020, Muralidharan and Sundararaman, 2011]. Across the world, the number of countries that use performance incentives for teachers doubled in the last decade, from one-third to two-thirds [The World Bank Group, 2018]. A large body of work has carefully measured the effect of performance pay for a fixed set of existing teachers. In a meta-analysis of teacher performance pay studies, there was substantial variation in effectiveness with an average increase in test scores of 0.09 SD [Pham et al., 2020]. In this paper, we seek to estimate whether there are sorting effects from performance pay in addition to direct behavioral effects.

1.3 A Model of Job Choice

The experimental design is motivated by a Roy [1951] model of job choice. First, we outline the worker's decision problem, in which they choose where to work. Then, given the employees' decisions, we demonstrate what types of employees firms will attract depending on the contract they offer.

Employee Job Choice

Employees choose between two jobs, j_F, which pays a fixed wage, w_0, or, j_P, which pays a wage dependent on the worker's output, y, and the piece rate, p. Output under performance pay is simply teacher's average output under a flat pay wage ("ability") , θ_i, plus their effort response to a performance pay contract ("behavioral effect"), β_i. Both are normally distributed with mean, μ_θ and μ_β, and variance, σ^2_θ and σ^2_β, respectively, and covariance $\rho_{\theta,\beta}$.

The wage from each contract is then:

$$w(\theta_i, \beta_i, j) = \begin{cases} w_0 & if\ j = j_F \\ py_i = p(\theta_i + \beta_i) & if\ j = j_P \end{cases} \tag{1.1}$$

Individuals do not have perfect information about their θ_i or β_i, so they make their job choice given their priors about these parameters. Their priors are a noisy function of the truth, $\hat{\theta}_i = \theta_i + e_i$ and $\hat{\beta}_i = \beta_i + \phi_i$, where $e_i \sim \mathcal{N}(0, \sigma_e^2)$ and $\phi_i \sim \mathcal{N}(0, \sigma_\phi^2)$. $\alpha_\theta = \frac{Var(E[\theta_i|\hat{\theta}_i])}{\sigma_\theta^2}$ and $\alpha_\beta = \frac{Var(E[\beta_i|\hat{\beta}_i])}{\sigma_\beta^2}$ capture teacher accuracy. An $\alpha_{\theta,\beta}$ of 1 is perfect information about their ability or behavioral effect, and an $\alpha_{\theta,\beta}$ of 0 implies that teachers have no information about their own ability or behavioral effect.

Jobs also carry non-wage utility, $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\mu^2)$, that is employee, i, and job, j, specific. These idiosyncratic tastes may include factors like commute time or firm amenities. Employees under performance pay also have dis-utility from the cost of additional effort they exert under these incentives.[1] Employees may also gain non-wage utility from the type of contract they receive, such as disliking inequality or enjoying competition. However, in section 1.8, we show that these preferences are not correlated with θ or β, so we exclude them from the model. An individual's total predicted utility is a linear combination of the wage and non-wage utility:

$$\hat{u}(\hat{\theta}_i, \hat{\beta}_i, j, \epsilon_{ij}) = \begin{cases} w_0 + \epsilon_{iF} & if\ j = j_F \\ p(\hat{\theta}_i + \hat{\beta}_i) - \frac{p\hat{\beta}_i}{2} + \epsilon_{iP} & if\ j = j_P \end{cases} \tag{1.2}$$

We will define the difference in predicted utility from performance pay versus flat pay as:

$$b_i = p(\hat{\theta}_i + \hat{\beta}_i) - \frac{p\hat{\beta}_i}{2} + \epsilon_{iP} - (w_0 + \epsilon_{iF}) \tag{1.3}$$

Therefore $b_i \geq 0$ implies the worker chooses a performance pay job.

Employee Quality by Job Type

We treat employment as a one-sided job choice by the employee. Employers accept anyone that applies to the firm.[2] However, employers can choose what contract they offer–a

[1] We assume employees exert effort, θ, under fixed pay which is determined based on their intrinsic motivation or career concerns. We assume employees have a quadratic cost of effort over additional effort exerted under performance pay. Therefore, the optimal additional effort under incentives is $\frac{p}{2c_i} = \hat{\beta}_i$, where c_i is the cost of effort parameter. The total cost of effort then is $c_i e^2 = c_i(\frac{p}{2c})^2 = \frac{p\hat{\beta}_i}{2}$.

[2] Section 1.5 will show this is a reasonable assumption in our setting. We will also relax this constraint by presenting results controlling for principal information to mimic settings where principals can screen employees.

flat pay contract or performance pay contract. The average output per worker, $\bar{y}(j)$, by contract offered is:

$$\bar{y}(j) = \begin{cases} E[\theta_i | b_i < 0] & if\ j = j_F \\ E[\theta_i + \beta_i | b_i \geq 0] & if\ j = j_P \end{cases} \tag{1.4}$$

Average output per worker at flat pay firms is the average employee ability for the subset of employees who choose flat pay ($b < 0$). Firms that offer performance pay receive both the average ability plus the effort response to performance pay, β, for the subset of teachers who chose performance pay ($b \geq 0$).

The difference in average output for firms that offer performance pay versus flat pay then is:[3]

$$\Delta\bar{y} = E[\theta_i + \beta_i | b_i \geq 0] - E[\theta_i | b_i < 0] \tag{1.5}$$

$$= \underbrace{E[\theta_i | b_i \geq 0] - E[\theta_i | b_i < 0]}_{\text{sorting on ability}} + \underbrace{(E[\beta_i | b_i \geq 0] - E[\beta_i | b_i < 0])P(b_i < 0)}_{\text{sorting on behavioral effects}} + \underbrace{E[\beta_i]}_{\text{avg. behavioral effect}} \tag{1.6}$$

The first term, "sorting on ability", captures the difference in average underlying ability between those who choose performance pay versus those who do not. The second term, "sorting on behavioral effect" represents the difference in behavioral response to incentives for those who choose performance pay versus flat pay. Together these two terms comprise the sorting effect of performance pay contracts, which together we will refer to as Δy_s. The last term ("average behavioral effect") captures the average behavioral response to performance pay for all teachers. This term is the effect of performance pay contracts on the static population of teachers, similar to what other studies of performance pay have focused on. Our focus for this paper will be to estimate both the sorting effects (the first two terms) and the direct behavioral effects (last term).

Model Predictions

The key predictions of the model are the existence of positive sorting in response to performance pay and the sensitivity of this positive sorting to teacher information and preferences.

If employees have any information about type (α_θ and/or $\alpha_\beta > 0$):
Prediction 1). Then $\Delta y_S > 0$: Performance pay induces positive sorting.
Prediction 2). $\frac{\partial \Delta y_S}{\partial \alpha_\theta} > 0$: Higher accuracy about type increases positive sorting.
Prediction 3). $\frac{\partial \Delta y_S}{\partial \sigma_\epsilon^2} < 0$: Higher variance in non-wage utility decreases positive sorting

To test each of these predictions, we conduct a randomized controlled trial. A key assumption of the model is that non-wage utility from a job is independent of the contract.

[3] Proof in A.3.

In our experiment, that assumption is satisfied by randomizing performance versus flat pay contracts across schools, allowing us to test predictions 1. In addition, we exogenously vary teachers' information about their ability via an information treatment and the variance of non-wage utility by varying the distance between jobs with opposite contract treatments, allowing us to test predictions 2 and 3.

1.4 Experimental Design

Timeline

Our design consists of two main phases: (i) the contract choice, where teachers are given the opportunity to choose their contract for the following year, and (ii) the randomized controlled trial, which randomizes schools to performance or flat pay contracts. The study was conducted from October 2017 to June 2019 with a private school chain that operates nearly 300 schools located across Pakistan. Figure 1.1 presents the timeline of interventions and data collection activities.

Phase 1: Contract Choice To understand whether higher-performing teachers prefer performance pay, we conduct a contract choice exercise with 2,480 teachers. Teachers were asked to choose between several contracts for the following year and told that the contract they chose would be implemented with some probability. The implied likelihood from the survey was that there would be a one-third chance their choice would be implemented.[4] Teachers were asked about two sets of choices: i). flat raise contract versus performance raise contract based on an objective measure of performance (percentile value-added), ii), flat raise versus performance raise based on a subjective measure of performance (principal evaluation).

We did several things during the implementation to ensure teachers understood this was a real, high-stakes decision. Two weeks before the survey, teachers received a description of the contract options they would be choosing between. During the survey itself, enumerators explained the stakes associated with the decision and showed teachers a video explaining the contract features and how their decision would be implemented with one-third chance. Teachers had to pass understanding checks before they were allowed to make the contract choice. We also played a coin flip game that we paid out in real-time to build trust in the survey. Finally, teachers in this system have previously experienced some forms of performance raises, though different from those conducted during the study, so they are familiar with some of the key aspects of these contracts.

Phase 2: Contract Randomization To measure the behavioral effects of performance pay, we randomize contracts across the remaining 243 schools that were not selected to

[4]Appendix figure A.14 presents information about how this probability was explained to participants, including screen captures from the video shown to participants. The actual implementation probability was a bit lower than one-third due to implementation constraints.

implement the teacher's contract choice. Schools were randomized to receive one of three contracts that determine the size of teachers' raises at the end of the calendar year.[5,6] The three contracts were:

- **Control: Flat Raise** - Teachers receive a flat raise of 5% of their base salary.
- **Treatment: Performance Raise** - Teachers receive a raise from 0-10% based on their within-school performance ranking.[7]

Performance Group	Within-School Percentile	Raise amount
Significantly above-average	91-100th	10%
Above-average	61-90th	7%
Average	16-60th	5%
Below average	3-15th	2%
Significantly below average	0-2nd	0%

There are two treatment sub-arms, which vary the performance measure used to evaluate teachers. Teachers are ranked within their school on either:[8]

[5]Triplet-wise randomization by baseline test performance was used, which generally performs better than stratification for smaller samples [Bruhn and McKenzie, 2009].

[6]To ensure teachers fully understood their contract, we conducted an intensive information campaign with schools. First, the research team had an in-person meeting with each principal, explaining the contract assigned to their school. Second, the school system's HR department conducted in-person presentations once a term at each school to explain the contract. Third, teachers received frequent email contact from school system staff, reminding them about the contract, and half-way through the year, teachers were provided midterm information about their rank based on the first six months. An example midterm information note is provided in appendix figure A.16. Control teachers were also provided information about their performance in one of the two metrics, in order to hold the provision of performance feedback constant across all teachers.

[7]Because the performance raise is a within-school tournament, this could potentially dissuade some high-quality teachers from sorting who would have otherwise if the incentive was absolute rather than relative. For example, if teachers believe all the best teachers will move into performance pay schools in the following year, then slightly above average teachers may choose not to sort because they would be a low performer relative to all of the very best teachers who are now at performance pay schools. However, we do not find evidence of teachers making this sort of assumption. When asked about the average change in quality in performance versus flat pay schools, teachers assumed performance pay schools would see an increase in average value-added of 0.006 SD. A difference of this magnitude would only dissuade positive sorting for those between the 50th and 51st percentile of the value-added distribution. Even if teachers could predict the actual level of sorting we find (0.013 SD), this should only dissuade teachers between the 50th and 52nd percentile from sorting. These effects would be minuscule in the scope of this experiment.

[8]The subjective and objective treatment arms have most features in common. Both treatments are within-school tournaments, so this holds the level of competition fixed between the two treatments. In addition, the variance in the distribution of the incentive pay is equivalent across the two treatments. The performance evaluation timeline also played out the same for all groups. Before the start of the year, managers set performance goals for their teachers irrespective of treatment. Teachers were evaluated based on their performance in January through December, with testing conducted in June and January to capture student learning in each term of the year.

- ***Objective Performance:*** Percentile value-added [Barlevy and Neal, 2012] averaged across all students they taught during the spring and fall term.[9]
- ***Subjective Performance:*** Principal evaluation at the end of the calendar year. Principals had discretion over how they would evaluate teachers but were required to communicate these criteria at the beginning of the year.[10]

We will present pooled results for subjective and objective incentives together for most results, unless their is a statistically significant difference between the two sub-arms. Along all of our main sorting outcomes, we cannot reject equality of effects between the two sub-arms. Understanding differences between the objective versus subjective treatment on teacher behavior is the focus of a companion paper [Andrabi and Brown, 2020].

The contract applied to all core teachers (those teaching Math, Science, English, Urdu, and Social Studies) in grades 4-13. Elective teachers and those teaching younger grades received the status quo contract. All three contracts have equivalent budgetary implications for the school. We over-sampled the number of subjective treatment arm schools due to partner requests, so the ratio of schools is 4:1:1 for subjective treatment, objective treatment, and control, respectively.

After schools have been assigned to different contracts, we then observe where teachers choose to work in the following year. Administrative data from the school system records which school a teacher is employed within the system or if they leave the school system.

Data

We draw on data from (i). the school system's administrative records, (ii). baseline and endline surveys conducted with teachers and principals (iii). endline student tests and surveys, and (iv). detailed classroom observation data.

Administrative data The administrative data details employee job description, salary, performance review score, attendance, and demographics for July 2015 to June 2019. It includes classes and subjects taught for all teachers, and end of term standardized exam scores for all students (linked to teachers).

Teacher and principal survey In addition to the contract choice exercise, the baseline survey included incentivized measures of teacher's beliefs about their performance along the objective (percentile value-added) and subjective (principal evaluation) metric. We

[9]Percentile value-added is constructed by calculating students' baseline percentile within the entire school system and then ranking their endline score relative to all other students who were in the same baseline percentile. Percentile value-added has several advantageous theoretical properties [Barlevy and Neal, 2012] and is also more straightforward to explain to teachers than more complicated calculations of value-added.

[10]These included items such as improving their behavioral management of students, assisting with administrative tasks, helping plan an after-school event, and improving students' spoken English proficiency. An example set of criteria are provided in appendix figure A.15.

also measured teachers' risk preferences using a high-stakes (a week's wage) and medium-stakes (half a day's wage) coin flip game and pro-sociality using responses to a volunteer opportunity. 40% of schools were randomly selected to participate in the baseline survey (and contract choice exercise). Data collection was conducted in October 2017, three months before the announcements of treatments.

At endline, we again measure teacher beliefs about their value-added, risk preferences, and offer a medium-stakes contract choice exercise. The survey also included measures of intrinsic motivation [Ashraf et al., 2020], efficacy [Burrell, 1994], and checks on what teachers understood about their assigned contract. The endline survey was conducted online with teachers and managers in spring and summer 2019. Appendix table A.15 lists the survey items used for each area along with their source.

The manager baseline and endline survey measured managers' beliefs about teacher quality, and the endline measured management quality using the World Management Survey school questionnaire.[11]

Endline Student Testing and Survey: An endline test was conducted in January to measure performance in Reading (English and Urdu), Math, Science, and Economics in grades 4-13.[12] The items were written in partnership with the school system's curriculum and testing department to ensure the appropriateness of question items. The research team conducted the grading. Items from international standardized tests (TIMSS and PERL) and a locally used standardized test (LEAPS) were also included to benchmark student performance. Students also completed a survey to measure four areas of socio-emotional development chosen based on the school system's student development priorities.[13]

Classroom Observation Data: To measure teacher behavior in the classroom, we recorded 6,800 hours of classroom footage and reviewed it using the Classroom Assessment Scoring System, CLASS [Pianta et al., 2012], which measures teacher pedagogy across

[11]Due to budget constraints, we were unable to have the World Management Survey research team conduct the survey. Instead, we asked managers to rate themselves on the rubric. This approach could result in inflated management scores. As a result, we use additional objective data to corroborate the management scores.

[12]The endline student test data was used both for evaluating the effect of the treatments and used to compute objective treatment teachers' raises.

[13]The areas are (i). love of learning (items drawn from National Student Survey, Learning and Study Strategies Inventory), (ii). ethical (items from Eisenberg's Child-Report Sympathy Scale, Bryant's Index of Empathy Measurement), (iii.) global citizen (items from Afrobarometer; World Values Survey), and (iv.) inquisitive (items from Learning and Study Strategies Inventory; Epistemic Curiosity Questionnaire). Appendix table A.14 lists the survey items used for each area along with their source. These are the four socio-emotional development areas they expect their teachers to focus on. These areas are posted on the walls in schools, and teachers receive professional development in these areas. Some principals also specifically make these areas part of teachers' evaluation criteria. In addition to four areas, the survey asked whether students liked their school.

a dozen dimensions.[14,15] We also recorded whether teachers conducted any sort of test preparation activity and the language fluency of teachers and students.

Measuring Teacher Ability

To measure teacher's "ability", θ, we calculate teacher value-added (VA) using student test scores from June 2016 and 2017, the two years prior to the randomized controlled trial. This allows us to measure teacher effectiveness in the absence of the treatments. We follow Kane and Staiger [2008] in constructing empirical Bayes estimates of teacher value-added. Teacher value-added is estimated as the teacher effect, μ, from a student-level equation:

$$y_{ijkcst} = \beta_0 + \sum_s \beta_s y_{ijkcs,t\text{-}1} \mathbb{1}[\textit{subject-grade} = s] + \sum_s \alpha_s y_{ijkcs,t\text{-}2} \mathbb{1}[\textit{subject-grade} = s] \quad (1.7)$$
$$+ \sum_s \gamma_s \bar{y}_{\text{-}ijkcs,t\text{-}1} \mathbb{1}[\textit{subject-grade} = s] + \chi_{st} + \psi_k + v_{ijkcst}$$
$$\textit{where } v_{ijkcst} = \mu_j + \theta_{ct} + \epsilon_{ijkcst} \quad (1.8)$$

where y_{ijkcst} is the test score for child i with teacher j at school k in class c in subject-grade s in year t. We regress these test scores on the student's one-year, $y_{ijkcs,t\text{-}1}$, and two-year, $y_{ijkcs,t\text{-}2}$, lagged test score in the given subject and the class's average lagged test score, $\bar{y}_{\text{-}ijkcs,t\text{-}1}$. We allow the coefficients on lagged test scores (β_s, α_s and γ_s) to vary across subject-grade. χ_{st} captures subject-grade-year shocks. ψ_k captures school-specific shocks. The residual, v_{ijkcst}, is the combination of teacher effects μ_j, classroom effects, θ_{ct}, and student-time specific shocks, ϵ_{ijkcst}. To isolate the teacher component, we use the residuals, v_{ijkcst}, to construct an empirical Bayes estimate of teacher value-added. We compute the average weighted residual and shrink by the signal variance to total variance ratio [Kane and Staiger, 2008].[16] Teachers for which we have few student observations are shrunk toward the mean teacher value-added (normalized to be zero).[17]

[14]There are tradeoffs between conducting in-person observations versus recording the classroom and reviewing the footage. Video-taping was chosen based on pilot data, which showed that video-taping was less intrusive than human observation (and hence preferred by teachers). Video-taping was also significantly less expensive and allowed for ongoing measurement of inter-rater reliability (IRR).

[15]We did not hire the Teachstone staff to conduct official CLASS observations as it was cost-prohibitive, and we required video reviewers to have Urdu fluency. Instead, we used the CLASS training manual and videos to conduct an intensive training with a set of local post-graduate enumerators. The training was conducted over three weeks by Christina Brown and a member of the CERP staff. Before enumerators could begin reviewing data, they were required to achieve an IRR of 0.7 with the practice data. 10% of videos were also double reviewed to ensure a high level of IRR throughout the review process. We have a high degree of confidence in the internal reliability of the classroom observation data, but because this was not conducted by the Teachstone staff, we caution against comparing these CLASS scores to CLASS data from other studies.

[16]VA is calculated as $VA_j = (\sum_t \frac{\bar{v}_{jt} h_{jt}}{\sum_t h_{jt}})(\frac{\hat{\sigma}^2_\mu}{\hat{\sigma}^2_\mu + (\sum_t h_{jt})^{-1}})$ where $h_{jt} = \frac{1}{Var(\bar{v}_{jt}|\mu_j)}$ and $\hat{\sigma}^2_\mu = Cov(\bar{v}_{jt}, \bar{v}_{jt-1})$. The first component of VA is the class-size weighted average class residual, and the second component is the shrinkage factor.

[17]Some of the classic problems with calculating VA (small classrooms, only observing the teacher with a

Having a teacher with a 1 SD higher VA for one year is associated with a 0.15 SD higher student test score. The effects are slightly larger for math, English, and Urdu and smaller for science. These effects are similar to other estimates from South Asia (0.19 SD, Azam and Kingdon [2014] and 0.15 SD, Bau and Das [2020]). Figure 1.2 shows the distribution of teacher value-added for the 3,687 teachers who teach in the school system at baseline.

Sample and Intervention Fidelity

Teacher and Principal Sample The study was conducted with a large, high fee private school system in Pakistan. The student body is from an upper middle-class and upper-class background. School fees are $900 USD. Table 1.1, panel A, presents summary statistics for our sample teachers compared to a representative sample of teachers in Punjab, Pakistan [Bau and Das, 2020]. Our sample is mostly female (81%), young (35 years on average), and the median experience level is 10 years, but a quarter of teachers are in their first year teaching. Nearly all teachers have a BA, and 68% have some post-BA credential or degree. Teachers are generally younger and less experienced than their counterparts in public schools, though they have more education. Salaries are, on average, $4,000 USD. Yearly turnover is 29%. There is a mix of career teachers and those who are less attached to their school. 70% and 36% expect to still be teaching at their current school in 1 year and 10 years, respectively. Panel B presents information about sample schools and principals compared to a representative sample of schools in India (data was unavailable for Pakistan) [Bloom et al., 2015]. Principals in our sample are more likely to be female and have much higher personnel management, operations, and performance monitoring scores than the average school in India.

Balance, Attrition, and Implementation Checks In this section, we provide evidence to help assuage any concerns about the implementation of the experiment. First, we show balance in baseline covariates. Then, we present information on the attrition rates. Finally, we show teachers and managers have a strong understanding of the incentive schemes. Combined, this evidence suggests the experiment was implemented correctly.

Schools in the two treatment arms and control appear to be balanced along baseline covariates. Appendix table A.5 compares schools along numerous student and teacher baseline characteristics. Of 27 tests, one is statistically significant at the 10% level, and one is statistically significant at the 5% level, no more than we would expect by random chance. Results control for these few unbalanced variables.

Administrative data is available for all teachers and students who stay employed or enrolled during the year of the intervention. During this time, 23% of teachers leave the school system, which is very similar to the historical turnover rate. 88% of employed teachers

single class of students, only one teacher per grade, infrequent student testing) are less of a concern in this setting. In our sample of grade 4-13 teachers, beginning in grade 6, teachers specialize and teach multiple sections of the same subject. On average, we observe 181 students across 5.6 classrooms per teacher over the two years of data. Schools are also relatively large, with an average of 131 students per grade. Students are tested every year, beginning in 4th grade.

completed the endline survey. While teachers were frequently reminded and encouraged to complete the survey, some chose not to. We do not see differences in these rates by treatment.

Finally, for the endline test, parents were allowed to opt-out of having their children tested. Student attrition on the endline test was 13%, with 3 pp of that coming from students absent from school on the day of the test and the remaining 10 pp coming from parents choosing to have students opt out of the exam. On both the endline testing and endline survey, we do not find differences in the attrition rate by treatment. We also do not find that lower-performing students were more likely to opt-out.

Teachers appear to understand their treatment assignment. Six months after the end of the intervention, we asked teachers to explain the key features of their treatment assignment. 60% of teachers could identify the key features of their raise treatment. Finally, most teachers stated that they came to fully understand what was expected of them in their given treatment within four months of the beginning of the information campaign. Knowledge of treatments in other schools is relatively low, though, which could impede sorting across schools. 15% of teachers could name spontaneously a school which was assigned to a given treatment arm.

1.5 Positive Sorting

We now present the main results of the paper in sections 1.5 through 1.8. In this section, we present evidence on Prediction 1. We first show that higher value-added teachers are more likely to choose performance pay contracts compared to flat pay when they are allowed to select their contract for the following year. We then show higher value-added teachers are more likely to move into performance pay schools after contracts have been randomized across schools. Finally, we document larger direct treatment effects for teachers who chose performance pay.

Positive Sorting on Ability

Measuring Contract Choices To measure teachers' preferences over contracts, we conduct a high-stakes choice exercise at baseline, where teachers' choice of contract is implemented with some probability. The survey states:

> *We can think of a raise as being a combination of two parts: the "flat" part that everyone gets regardless of their [subjective/objective] score and the "performance" part where those with higher [subjective/objective] scores receive more than those with low [subjective/objective] scores. What percentage of the raise would you like to be flat?".*[18]

We ask this question twice: once for an objective performance metric (percentile value-added) and once for a subjective performance metric (principal evaluation). A.4 provides

[18] As a robustness check, we also ask the question in a simpler way. We ask teachers to choose between five options, from a completely flat up through a completely performance-based raise. 76% of teachers give an internally consistent answer across the two versions of the question.

the full question description, including the examples given, understanding checks preceding the question, and explanation to teachers about how percentile value-added is calculated.

Figure A.4 shows the distribution of teachers' responses. Most teachers want at least part of their raise to be performance-based, with less than 10 choosing a completely flat raise. On average, teachers wanted 56% of their raise to be performance-based when the performance metric was subjective and a slightly lower 52% when the performance metric was objective. For ease of communication going forward, we will group responses that are greater than 50% flat as "chose flat pay" and less than or equal to 50% as "chose performance pay". As an alternative, the appendix presents results treating the choice as a continuous variable. All of the main results are unchanged between the two approaches.

Figure 1.3 presents the relationship between contract choice and teacher demographics, characteristics, and beliefs. A strong predictor of contract choice is the teacher's belief of their principal's rating of them in the next year. Teachers that are more risk-loving (as measured in a real-stakes coin flip game) and those that say they are likely to stay teachers over the next five years also prefer performance pay. Female teachers are less likely to choose performance pay, and experienced teachers are slightly more likely to choose performance pay. These relationships generally hold whether the performance metric is subjective or objective (shown in Figure A.2).

Positive Sorting in Contract Choice We find that teachers who chose a performance pay contract have significantly higher baseline value-added. Figure 1.4 plots the distribution of baseline value-added (in student standard deviations) for teachers who chose performance pay (solid line) versus those who chose flat pay (dashed line). The entire distribution is shifted to the right for those who wanted performance pay, and the difference is equivalent to a 0.05 SD difference in test scores. This difference holds for the choice between objective performance pay versus flat pay and subjective performance pay versus flat pay.

To test whether there is a significant difference in value-added by contract choice we estimate:

$$VA_{i,t-1} = \beta_0 + \beta_1 ChosePerfPay_i + \epsilon_i \tag{1.9}$$

where $VA_{i,t-1}$ is a teacher's baseline value-added (our measure of teacher quality in the absence of incentives), and $ChosePP_i$ is the contract the teacher chose at baseline. Throughout the results section, $ChosePP_i$, refers to their baseline survey choice, *not* the contract teachers actually received.

Table 1.2 presents the results from eq. 1.9. As we showed in the figures, teachers who chose performance pay had 0.05 standard deviation higher baseline value-added. The relationship is similar whether we look at choices on objective or subjective performance pay. Columns (2) and (4) control for the principal's evaluation of the teacher. We see that principals do have some information about teacher value-added. A 1 SD increase in principal rating is related to a 0.02 SD increase in value-added. However, when we control for the information that principals have, the teacher's choice of performance pay is still a significant predictor of value-added. This suggests that teachers have additional information about their own quality beyond what principals know.

While on average teachers seem to have information about their ability, we do see heterogeneity across teacher type. Figure 1.5 presents the relationship between baseline value-added and likelihood of choosing performance pay by teacher gender, age, and experience. Here a steeper line suggests more positive sorting in response to performance pay. The average level of the line shows the extent to which performance pay is preferred on average for that sub-group. First, we see female teachers are less likely in general to prefer performance pay but have a similar relationship between ability and contract choice as male teachers. We also see that more novice teachers appear to have less information about their ability or, at least, are not sorting on that information. However, we also see that older teachers may be more overconfident and their abilities and, therefore, more likely to choose performance pay even when they are not actually high ability.

Measuring Job Choice Next, we investigate whether the composition of teachers changes between flat pay versus performance pay schools. We use administrative data from the school system to identify where each individual works at baseline (December 2017) and a year after the contracts are announced (December 2018). We observe if a teacher joins or leaves the school system but do not know if and where they are employed if they leave the school system.[19,20] During the treatment information campaign, teachers were also told if they transferred schools, they would be subject to the contract of the school they transferred to.[21] Transfers are initiated by the teacher and need to be accepted by the receiving school.[22] Transfers are nearly always accepted by the receiving school. This is because incumbent teachers have hiring priority, and there is high turnover within the system, virtually guaranteeing open positions at the school of interest each summer. Therefore it is appropriate to think of this setting as a one-sided choice problem, as the schools have little say in who within the transfer applicants is hired.

[19]We also can see whether teacher's actual job choice is correlated with their contract choice. As we would expect, teachers who chose performance pay at baseline are more likely to move into performance pay schools. This serves as a helpful check on the consistency between our contract choice and job choice outcomes.

[20]There is substantial churn throughout the system. Transfers across schools are common (15% of teachers), and turnover is high (23%).

[21]Teachers were provided information about other schools' treatment status over email and through their employee portal. This ensured full information for all study participants, allowing the possibility of positive selection. Teachers were also reminded of their school and other schools' treatment status during the summer break via email and their employee portal, as that is the time most transfers take place.

[22]There are two types of transfers. Many schools operate on a larger campus. For example, there may be a primary school, middle school, and high school all on the same larger campus, and a teacher applies to transfer from the primary school to the middle school. For example, the other type is across campuses transferring from a middle school teacher at a school in Lahore to a different branch of the school system in Karachi. 6% of teachers make a within campus transfer, and 11% of teachers make an across campus transfer each year. Transfers are recorded in the administrative data, and we can observe rejected transfer applications. The vast majority of transfers and resignations happen over the summer break between school years (calendar of transfers shown in figure A.3).

Positive Sorting in Job Choice Figure 1.6 presents the distribution of teacher value-added at baseline (Panel A) and then one year after the announcement of the contract (Panel B) across treatment and control schools. At baseline, the two distributions are virtually indistinguishable. However, a year later, there are now more below-average value-added teachers in flat pay schools and more above-average value-added teachers in performance pay schools, with an average difference of 0.022 SD. Similarly, we can see the cumulative distribution functions lie on top of each other at baseline, but, a year later, the performance pay schools dominate flat pay schools at every part of the distribution (figure A.8).

To test this formally, we estimate the quality of individuals who end up in performance pay schools after a year:

$$VA_{i,t-1} = \beta_0 + \beta_1 WorkatPP_i + \beta_2 Post_i + \beta_3 WorkatPP_i * Post_i + \chi_j + \epsilon_i \tag{1.10}$$

$WorkatPP$ is a dummy for whether a teacher works at a school assigned performance pay, $Post$ is a dummy, which is 1 for December 2018, the end of the intervention, and 0 for December 2017, the month before the announcement of treatments. We control for randomization strata and cluster standard errors at the level of school (the unit of randomization). β_1 tells us the difference in quality between schools assigned performance raises versus flat raises just before the treatments were announced. This coefficient is a test of balance between the treatment and control schools, as there should be no difference in teacher quality at baseline. β_2 tells us the change in the quality of teachers teaching at flat pay schools between the beginning and end of the intervention year. β_3 is the key coefficient of interest. It tells us whether performance pay schools attracted better teachers over the year of the intervention relative to flat pay schools.

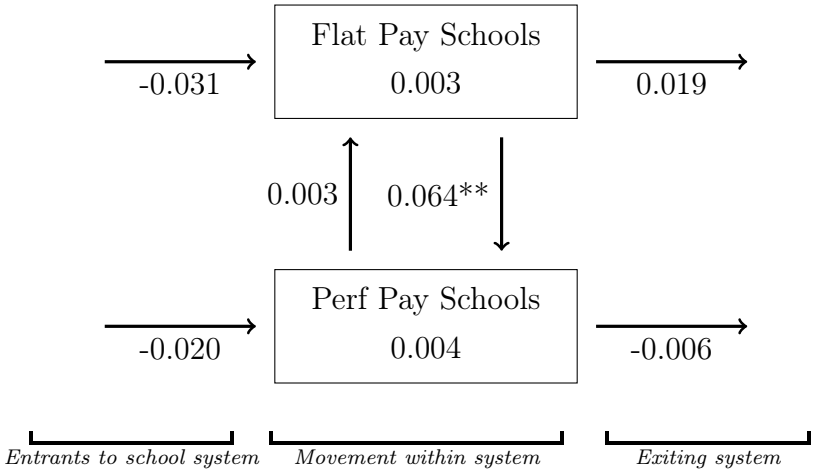
Table 1.4, column 1, presents the results of eq. 1.10. As we saw with the figures, there is no difference between performance and flat pay schools at baseline. However, a year later, the average baseline value-added of teachers at flat pay schools is 0.019 SD lower in flat pay schools and 0.003 SD higher in performance pay schools (a difference of 0.022 SD between treated and control schools). The magnitude of this effect is relatively small, but as this was just a one-year contract change, it is not surprising we do not find huge shifts in employment across schools. As this is the extent of positive sorting from a one-year contract change, we would expect this to be a lower-bound on the extent of sorting.

The results are robust to additional controls in columns 2 and 3 for region, grade, and subject. Column 4 adds controls for the principal's rating of the teacher. Principals appear to have some information about teacher quality. A 1 SD increase in the principal's rating of the teacher is associated with a 0.13 SD higher teacher value-added (0.02 SD in student standard deviations). However, the coefficient on $WorkatPP_i * Post_i$ remains significant when we control for principal information, so this sorting behavior is providing a signal about teacher's quality beyond what principals know already, suggesting teachers do have private information. We do not see any significant differences in sorting by gender, age, or experience.

Switchers, Leavers, and New Entrants The job choice results we have shown could come from two sources of self-selection: teachers switching within the system (going from a

flat pay school to a performance pay school or vice versa) or teachers differentially leaving the school system from flat versus performance pay schools. Until this point, we have not included any results on new entrants into the school system that started working during the intervention or the semester before because we do not have a measure of value-added for them prior to the intervention. For teachers who entered during the interventions, we can calculate their value-added based on their student's June 2019 scores. The concern is that this could capture both innate teaching ability and treatment effect. However, the school system does not provide new teachers with any performance incentives during their first year, so the effect would come from a misunderstanding of their contract or from positive spillovers from other treated teachers.

The diagram below maps the change in teacher quality for teachers who switch within the system, leave the system, and are new entrants to the system during the intervention year. The numbers next to each arrow show the average baseline value-added for that group. For example, the arrow in the top left part of the diagram shows that the average value-added for teachers who are entering the school system and starting their first job at a flat pay school is -0.031 SD. The numbers inside the boxes show the average value-added for teachers who stayed at their original school or moved from a school to another school with the same treatment. For example, teachers who stayed at a flat pay school or moved from one flat pay school to another flay pay schools had an average baseline value-added of 0.003 SD.



We can see that most of the effect is driven by higher quality teachers leaving control schools and moving into treatment schools. The average value-added of those who moved from flat pay to performance pay schools is 0.064 SD. Whereas, the average quality of those who moved from performance pay to flat pay is 0.003 SD. We also see better teachers leave the school system from flat pay schools (0.019 SD) than performance pay schools (-0.006 SD), which is consistent with positive sorting, but the difference is not statistically significant. We do not see significant differences in the quality of teachers who stay at their current school or among new entrants. It is not surprising that we do not see effects among new entrants as the study was not set up to test this (see Leaver et al. [2019] for a test of this type of

sorting). The treatments were not advertised to new hires and were set to expire before new hires would begin receiving them.

Positive Sorting on Behavioral Effect

Do teachers who chose performance pay also have larger behavioral responses? To test prediction 1 for behavioral effects, we compare the treatment effect of performance pay for those that chose performance pay versus flat pay in the baseline survey:

$$\begin{aligned} TestScores_i = \beta_0 + \beta_1 AssignedPPtreat_j + \beta_2 ChosePP_i \\ + \beta_3 AssignedPPtreat_j \cdot ChosePP_i + \beta_4 TestScore_{i,t-1} + \chi_j + \epsilon_i \end{aligned} \tag{1.11}$$

The outcome is endline test scores for students taught by teacher, i. $PPtreat_j$ captures the treatment assigned to the teacher's school, j for the school at which the teacher taught at the time of treatment announcement. As we saw in section 1.5, some teachers change schools during the experiment, so $PPtreat_j$ gives us the intent-to-treat effects of performance pay. $ChosePP_i$ is the teacher's contract choice from the baseline survey. We control for randomization strata, χ_j, and student's baseline test scores, $TestScore_{i,t-1}$. Standard errors are clustered at the school level (the unit of randomization). The coefficient of interest is β_3, which captures whether there is a differential effect of performance pay on teachers who wanted that contract. We, of course, restrict to the RCT sample of schools, so the $ChosePP_i$ variable is unrelated to the contract assigned, $AssignedPPtreat_j$.

We find that teachers who wanted performance pay have much larger behavioral responses than those who wanted flat pay, (0.09 SD versus 0.01 SD). Figure 1.7 presents the average effect of performance pay across all teachers and then splits the sample by teachers who chose performance pay versus those who chose flat pay. Table 1.5, column 2, presents the results of equation 1.11. Column 3 controls for principal rating, which does not change our effects. In fact, along this metric we do not find that principals have information about teacher quality. Results, shown in table A.2, are also identical if we treat contract choice as a continuous variable (percent of raise chose to be performance-based).

Is this "sorting on behavioral effect" just picking up the same high value-added teachers who wanted performance pay? It does not appear that is the case. Column 4 shows there is no relationship between baseline value-added and behavioral effect. Column 4 shows that the coefficient on $PPtreat_j \cdot ChosePP_i$ remains stable when we control for value-added and value-added interacted with treatment. This suggests that high "ability" teachers and high "behavioral effect" teachers are not the same individuals.

Total Effect of Performance Pay Returning to our decomposition of the total effect of performance pay eq. 1.5, we have the following total effect:

Type of effect	Effect (student SD) Contract Choice	Job Choice
Total Sorting effect:	**0.074**	**0.033**
Sorting on ability	0.049	0.022
Sorting on behavioral effect	0.025	0.011
Behavioral effect:	**0.066**	**0.066**
Total	**0.140**	**0.099**

We summarize the effect of each of these components in the setting without switching costs (contract choice exercise) and with high switching costs (teacher job choice in the second year). When we incorporate sorting effects, we see that the total effect of performance pay is somewhere between 110% and 50% larger than measuring just the effect on the existing stock of teachers.

1.6 Asymmetric Information

How much information do employers have?

As we saw in table 1.2 and 1.4, principals do have some information about teacher quality. However, the extent of principal information varies substantially depending on the dimension of teacher quality and principal's exposure to teachers. At endline, we ask principals to rate teachers they oversee along four dimensions of quality: i). attendance, ii). managing student discipline in the classroom, iii). incorporating higher-order skills in lessons, such as analysis and inquiry, and iv). value-added. We then compare this to teachers' actual daily attendance, recorded via biometric clock in/out data, teachers' management of student discipline, and incorporation of higher-order skills assessed using classroom observation data, and teachers' actual value-added.

Table 1.6 presents the relationship between principals' beliefs and teachers' actual outcomes. Pooling across all four dimensions (column 1), we see principals are decently well-informed. A 1 SD increase in teacher outcome is associated with a 0.17 SD increase in principal rating. However, when we look at each dimension separately, we see principals do much better in rating criteria that are highly observable–teacher attendance and student discipline–which have a coefficient of 0.19 and 0.23, respectively. Along more subtle areas of teaching practice like developing analysis and inquiry skills and value-added, principals are much worse at predicting teacher quality (0.14 and -0.04, respectively). More experienced principals are not any more accurate in rating teachers (column 6).

We also find that principal accuracy varies substantially depending on the level and type of exposure principals have with teachers. From September 2018 to January 2019, we randomly assign some teachers to receive more frequent classroom observations from their principals. Principals were instructed to observe treated teachers at least once a month during the period, though not all principals completed the full set of observations. We find

that treated teachers receive 2.7 observations during the 5-month period, relative to 1.8 for the control.

Principals provide much more accurate ratings for teachers who were assigned to the observation treatment. Table 1.6, column 7, provides principal rating by observation treatment status. A 1 SD increase in teacher outcomes is associated with a 0.06 SD increase in principal rating for control teachers versus 0.25 SD for treated teachers. This increase in accuracy comes both from increasing their rating of high performers and lowering their rating of low performers.

However, principals actually get *less* accurate the longer they work with a teacher. Table 1.6, column 8, compares principal accuracy for principals who have worked at the same school as the teacher for more than or less than two years.[23] A 1 SD increase in teacher outcomes is associated with a 0.18 SD increase in principal rating for teachers whom they have overlapped with less than two years versus 0.01 SD for those they have overlapped with for more than two years.[24] These effects are driven by principals boosting scores of low performing teachers the longer they overlap with them (figure A.9).

Because overlap is not randomly assigned in this context, we cannot be sure if this effect is the causal effect of overlap or something correlated with it. For example, the amount of time overlapping would also correlate with principal experience and job change frequency. While we cannot address every possible omitted variable, column 9, controls for principal and teacher years of experience, and column 10 controls for principal fixed effects. Our results are robust to the addition of these controls.

How much more information do teachers have?

Much of the sorting value of performance pay schemes depends on how much more information teachers possess relative to their employers about their ability. To assess this, we compare the explanatory power of characteristics schools can observe (experience, age, and credentials) and principals' rating to using teacher's contract choice. Figure 1.8 plots predicted teacher value-added relative to actual value-added for each of these models. The solid line is from predicted value-added using age, experience, and credential-type fixed effects. We see that these criteria predict some variation in teacher value-added. The dashed line adds principal evaluation data to the model, which slightly improves the model (though we cannot reject equality of the two models). Finally, adding in teacher contract choice (dotted line) triples the predictive power of the model. This suggests that teachers have substantially more information about their type than their employer.

We find the extent of asymmetric information varies over a teacher's tenure. Figure 1.9 presents the coefficient on the regression of predicted value-added on actual value-added. The solid black circles and 95% confidence intervals show the coefficient when predicted

[23]Here "overlap" is just employment at the same school. This does not imply that the person who is currently the principal was the teacher's manager for the entire time. They may have worked together both as teachers or the principal may have previously been in another administrative role at the school that did not involve overseeing that teacher.

[24]Results are similar if we treat overlap as a continuous variable in years rather than a dummy.

value-added is constructed using just principal evaluation data. The gray diamonds show the coefficient when we add teacher contract choice to the prediction. The data is split by novice (less than 3 years), experienced (3-8 years), and very experienced teachers (greater than 8 years). We see an interesting pattern across teacher experience. As we showed in the effect of overlap with a teacher, principals become less accurate the more experienced a teacher is. Teachers initially become more accurate with experience but drop off for very experienced teachers. Teachers have more information than principals in all years except for very novice teachers.

What is the source of teacher's private information? There are two possible explanations for this result: (i) teachers have information about their own ability or (ii). teachers do not have information about their value-added, but value-added is correlated with other preferences (risk, competitiveness, etc.) that make high types more likely to choose performance pay. We do not find evidence for the second claim. Higher value-added teachers and those that have larger behavioral responses do not have different risk preferences, preferences for competition, or pro-sociality (table A.3). We can also control for risk preferences, preferences for competition, and pro-sociality in our main positive sorting results on ability and behavioral effect (table A.4). Our results remain unchanged when we control for these potential channels.

1.7 Magnitude of Positive Sorting

Our experiment allows us to explicitly test predictions 2 and 3, to see the effect of teacher's information and switching costs on the extent of positive sorting. First, we exploit randomization of the neighboring school's treatment as exogenous variation in switching costs. Second, we randomly provide some teachers with historical information about their performance to test the effect of private information.

Sorting by teacher information

Another potential driver of positive sorting is how accurate teachers are about their own ability or their behavioral response. To test whether teacher's information about their own performance affects positive sorting, we randomize teachers to receive information about their value-added from the prior year during the endline survey. A random subset of teachers received the following message during the survey before they made their contract choice. *Based on your students' test scores last year, you were in the [X] percentile. This means you performed better than [X] percent of teachers. You would have been in the [Y] appraisal category. In an average year, this would mean you'd receive a raise of [Z].*

First, for this information treatment to work, teachers must not be fully informed about their own value-added. We find that teachers update in response to this information treatment. Figure 1.11, panel A, plots teacher's predictions about their performance in the coming year relative to their true performance that year for teachers who received no information versus those who learned about their historical value-added. Those that

receive information do a better job of being able to predict their future value-added. This information also influences their ultimate contract choice. The correlation between choosing performance pay and teachers increases by 50% for those assigned to the information treatment versus no information, as we see in figure 1.11, panel B. This suggests that better information about one's own ability does increase the extent of positive sorting.

Sorting by switching costs

The extent of positive sorting may depend on how strong their preferences are for wage versus non-wage utility, such as location or firm amenities. We can explicitly test this prediction by comparing teachers who face different switching costs to achieve their desired contract. We do this by exploiting random variation in the treatment of a teacher's neighboring school.

Most schools operate on a larger campus, which contains multiple schools (primary school, middle school, high schools). Within the same campus, different schools may be assigned to different contracts. Therefore, we can look at the extent of positive sorting when another school on the same campus was assigned to the opposite treatment as the teacher's own school's treatment. For example, we can see that in one of the cities, Lahore, shown in appendix figure A.10, there are a mix of treatment and control assignments across schools within the same campus. We define the "closest school" as the school on the same campus as the teacher currently works, with grade levels closest to the teacher's current assignment. For example, for a first-grade primary school teacher, the "closest school" is the pre-primary school (nursery through kindergarten) on the same campus. However, for a fifth-grade primary school teacher, the "closest school" is the middle school (grades 6-8) on the same campus.

Our main specification is:

$$\begin{aligned} VA_{i,t-1} = \beta_0 &+ \beta_1 WorkatPP_i + \beta_2 Post_i + \beta_3 WorkatPP_i * Post_i + \beta_4 OppTreat_i \\ &+ \beta_5 OppTreat_i * Post + \beta_6 OppTreat_i * WorkatPP_i \\ &+ \beta_7 OppTreat_i * WorkatPP_i * Post + \chi_j + \epsilon_i \end{aligned} \tag{1.12}$$

This is similar to eq. 1.10 but adds in interaction with $OppTreat_i$, which is a dummy for whether the closest school is assigned the opposite treatment as the teacher's own school. The coefficient of interest is β_7, which tells us the difference in the extent of positive sorting for teachers who would face smaller switching costs to receive their ideal contract.

We find that when teachers' closest school is assigned the opposite treatment, there is a higher rate of positive sorting. Table 1.8 presents these results. Column 1 shows the extent of positive sorting for the full sample. Column 2 and 3 split the sample by whether the closest school received the same or the opposite treatment as the teacher's own school. The magnitude of positive sorting is about four times larger (0.04 SD versus 0.009 SD). Column 4 presents eq. 1.12. While there is a large difference in the extent of sorting, we cannot reject equality of the coefficients at the 10% level.

Another approach to test whether switching costs dampen the extent of positive sorting is to compare the contract choice versus the job choice in the second year. We can think of the contract choice decision as zero switching cost because teachers could remain at their current position but receive their preferred contract. Job choice decisions in the second year is a relatively high switching cost, as teachers move across schools in response to a short-term acquisition of their preferred contract. Comparing these two settings, we see substantial differences in the extent of positive sorting (0.05 SD versus 0.022 SD).

1.8 Potential Negative Consequences of Sorting

Does performance pay attract "cheating" teachers?

We have shown performance pay allows schools to attract "good" types along several dimensions, but we may be concerned that it also attracts teachers who know how to "cheat" the performance pay system. For example, it may attract teachers who are willing to change their teaching to maximize financial gain while sacrificing some areas of student development. To test for this type of negative sorting, we look at effects in three areas: i). teaching pedagogy (using classroom observation data), ii). student socio-emotional development (using a student survey) and iii). memorization behavior (as measured by performance across different question types at endline).

First, we do not find that teachers who prefer performance pay are more likely to engage in distortionary teaching practices. They are significantly less likely to exhibit these behaviors than teachers who did not want performance pay. Figure 1.13 and appendix table A.7 presents the treatment effects of objective performance pay along several dimensions of teaching pedagogy (classroom climate, differentiation, student-centered focus, and time spent on test preparation). The coefficient of interest is *Chose Perf Pay* Perf Pay Treat*, which tells us the heterogeneity in treatment effect by whether the teacher chose performance pay at baseline. The row titled $\beta(Treat + Treat * ChosePP)$ also presents the effect of performance pay for teachers who chose it. As we show in a companion paper [Andrabi and Brown, 2020], we find that objective performance pay results in a more negative classroom climate (more yelling, stricter discipline), more teacher-led time (less student-centered), and more time teaching to the test. However, these negative effects are almost completely concentrated among teachers who did not want performance pay. The overall effect of objective performance pay on classroom pedagogy rating is -0.41 SD for teachers who did not want performance pay as opposed to 0.16 SD for teachers who did want performance pay.

Second, we do not find that teachers who prefer performance pay ignore other areas of student development in order to maximize their pay. Figure 1.14 and appendix table A.8 present results. At endline, we measure student satisfaction and socio-emotional development along five dimensions (survey items shown in appendix table A.14). The effect of objective performance pay for teachers who chose flat pay is generally small and mixed across different dimensions. However, for teachers who chose performance pay, we find a significant positive

effect on three of the five areas with an overall effect of 0.12 SD.

Finally, we can zoom in on different question types from the endline exam to see if treatment effects are concentrated among memorization-type questions, at the cost of other knowledge and skills. Table A.6 column 1 presents the results for all question types. Column 2 presents results for questions that were pulled from external sources (PISA, TIMSS, and LEAPS), and hence were unlikely to be questions students would have been able to memorize. Columns 3 and 4 include questions from one grade below and one grade above the student's current year. We find significant effects of performance pay for teachers who chose it along all three areas, ranging from 0.11 SD to 0.20 SD. Combined, this evidence shows that the negative consequences that are often associated with performance pay are concentrated among teachers who did not want those contracts, not those who would sort in.

Does performance pay push out altruistic teachers?

Another concern is that performance pay may drive away teachers who are intrinsically motivated or pro-social. To test this, we measure teachers' pro-sociality, efficacy, competitiveness and time spent on school public goods (such as helping other teachers or assisting with extra-curriculars).[25] Figure 1.10 presents the difference along each characteristic for teachers who chose performance pay versus flat pay. We do not find that teachers who prefer performance pay spend significantly less time on providing public goods. Teachers who chose performance pay spend slightly more time on collaboration with other teachers and the same amount of time on administrative tasks. They do, however, spend less time meeting with parents and more time grading than those who chose flat pay. Teachers who prefer performance pay have similar levels of pro-sociality (as measured by signing up to volunteer to help financially disadvantaged students). They also are less likely to view their current job as a stepping stone to another job. This evidence suggests that performance pay does not attract significantly less altruistic teachers.

1.9 Policy Counterfactuals

In addition to understanding the extent of sorting when individual schools offer performance pay contracts, we may be interested in the effect of a whole school district or state introducing performance pay. It is also useful to understand the effect of introducing the policy for a longer period as we would expect sorting effects be much larger for permanent contract changes. To conduct these counterfactual exercises, we use estimates of teacher's priors, distribution of ability and behavioral response, and elasticity of supply to a given job from our experiment. We then simulate the effects of a longer term performance pay policy, applied to a larger set of schools.

[25]Survey item description and sources are presented in appendix table A.15. Most measures are based on teacher self-report, though, so we may be concerned about some response bias. It is not clear if this bias would be differential by contract choice.

We augment the simple framework from section 1.3 to make the employment decision a bit more realistic. First, workers now choose between many jobs, j, across the teaching and non-teaching sectors, with a cost, c, to change sectors. Employees make the decision of which job to work at in a given period based on: i). the expected flow of wages, w_{jt}, for their remaining time in the labor force, τ, ii). the cost to change sectors if the job is not in the sector the employee currently works in, iii). non-wage utility, which is employee-job (ϵ_{ij}) and employee-job-time (ϵ_{ijt}) specific. Flat pay jobs pay a wage of 0, and performance pay jobs pay the piece-rate, p, times workers' priors about their output ($\hat{\theta} + \hat{\beta}$). Whether a job offers performance pay in a given year is denoted by δ_{jt}. Employees have full information about what contracts will be provided by each job over the length of their time in the labor force.

Employees choose which job has the highest predicted utility:

$$u_t(\theta_i, \beta_i, j, \tau_i) = \max_j \left(\sum_{t=1}^{T} w_{jt} \mathbb{1}[\tau_i > t]\right) - c_i \mathbb{1}[s_t \neq s_{t-1}] + \epsilon_{ij} + \epsilon_{ijt}$$
$$\text{where } w_{jt} = [p(\hat{\theta}_i + \hat{\beta}_i]\mathbb{1}[\delta_{jt} = 1]$$

Table 1.9 presents the key parameter values used. To calculate the mean and standard deviation of teacher ability and behavioral effect of incentives, we make the following assumptions about the test score function. For the pre-period (and control group): $y_{it} = \theta_i + e_{it}$. For the treatment group during the intervention: $y_{it} = \theta_i + \beta_i + e_{it}$. We use our calculation of value-added in a given year for y_{it} and assume $Cov(e_{it}, e_{it+1}) = 0$. Here $t-1$ is one year before the intervention, t is the baseline and $t+1$ is the intervention year. The first and second moments of θ and β and their covariance are:

$$\begin{array}{ll} \mu_\theta = \bar{y}_{it} & \sigma^2_\theta = Cov(y_{it-1}, y_{it}) \\ \mu_\beta = \bar{y}^T_{it+1} - \bar{y}^C_{it+1} & \sigma^2_\beta = Var(y^T_{t+1}) - Var(y^T_t) - 2[Cov(y^T_{it}, y^T_{it+1}) - Cov(y_{it-1}, y_{it})] \\ & \rho_{\theta,\beta} = Cov(y^T_{it}, y^T_{it+1}) - Cov(y_{it-1}, y_{it}) \end{array}$$

Our estimates of σ^2_θ and σ^2_β come from the existing set of teachers in the school system. However, the distribution in quality in the entire labor force is likely larger, so we offer optimistic values of the these parameters as well.

The variation in job-employee specific non-wage utility comes from distribution of employee-job fixed effects from a regression of job choice on wage and fixed effects during the years before and during the policy. The variation in job-employee-time specific non-wage utility comes from the distribution of residuals from the same specification. The mean and variance in the cost to change professions comes from survey responses in the endline survey conducted with teachers.

Finally, the accuracy of teachers' priors about their ability, α_θ, and behavioral response, α_β for existing teachers come directly from the contract choice experiment. We use a separate set of lower accuracy, but non-zero, priors for individuals who are not currently teachers. The values chosen take into account evidence from this study across teacher tenure and evidence on applicant teacher accuracy from Leaver et al. [2019] and Johnston [2020]. We

also include optimistic values of the parameters to take into account that longer term policies would likely result in better understanding of the performance metrics used.

We find that the introduction of a long term performance pay contract induces a fair amount of sorting, though effects vary depending on the use of pessimistic versus optimistic parameter values. Figure 1.15 presents the effects over time of introducing a 1 year, 10 year or 30 year performance pay policy. The effect of a 1 year policy is just the average behavioral response (0.07 SD). Under a 10 year policy, there is an average effect of 0.075 SD (0.10 SD) during the time the policy is in place, if using pessimistic (optimistic) parameter values . Under optimistic parameters, there are also effects after the policy is removed due to the attraction of higher performing teachers that then stay in the profession even after the policy is removed. The introduction of a 30 year policy results in an average effect of 0.09 SD (0.17 SD) under pessimistic (optimistic) parameters. These effects are 1.3-2.4x larger than the one year effects of performance pay.

1.10 Conclusion

In this paper, we conduct a choice exercise and randomized controlled trial to understand whether performance pay allows schools to attract and retain better teachers. We find that teachers appear to have information about their ability (value-added) and behavioral response to incentives. Teachers who are higher ability and have larger behavioral responses significantly prefer performance pay. Teachers' contract choices are also significantly predictive of performance even controlling for the characteristics schools have access to, such as experience, credentials and performance evaluation scores. This suggests that there is asymmetric information between employees and employers about employee quality. We also find that performance pay does not attract teachers with unfavorable characteristics, such as those who contribute less to public goods or focus on maximizing their incentive pay at the cost of more well-rounded student development.

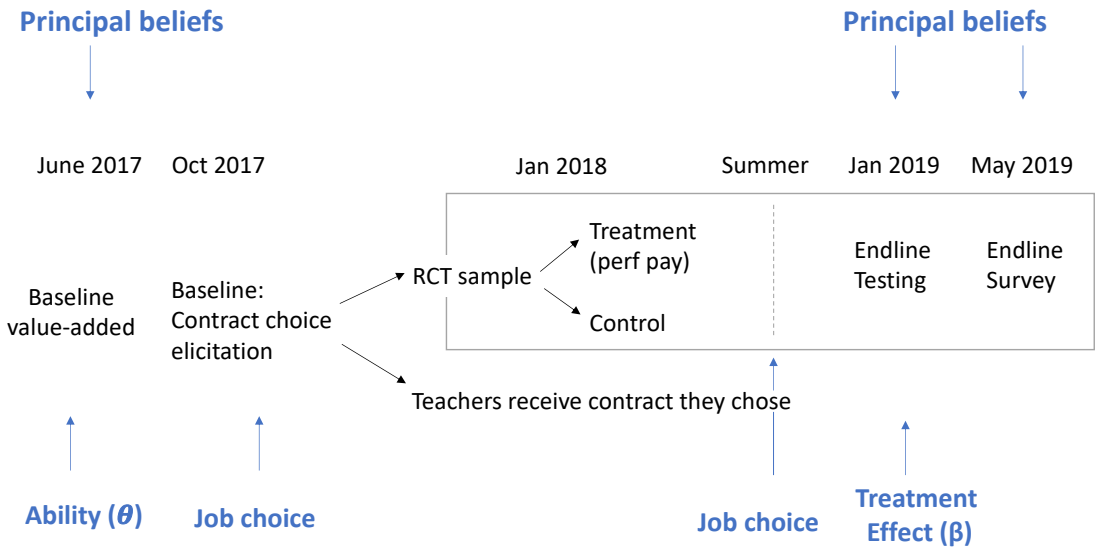
To understand what the effects of different policies would be on the extent of sorting, we use additional exogenous variation to test the effect of increasing teachers private information and lowering the switching cost to access their preferred contract. We find teachers are responsive to both of these margins and both increase the extent of sorting. Taking the results from the main experiment and the comparative static results, we are able to simulate policy counterfactuals. While the results are sensitive to the choice of parameter values, we find that the long term effects of performance pay are 1.3-2.4x the effects of a one year policy due to sorting.

One limitation of the study is the inability to look at long run effects directly in the experimental sample and having to rely on other papers to estimate the extent of private information that exists among those who are not currently in the teaching sector. Understanding the features of this population in an important area for further work. Another limitation is understanding where high quality potential teachers are drawn from as the social welfare implications of pulling high quality workers from other sectors varies substantially.

The implication of these findings is that firms should take advantage of information employees have to help improve the quality and match of their employees. We also see that increasing worker's autonomy to select the contract they prefer significantly improves firm and worker outcomes. Finally, the findings suggest that previous evidence on the effect of performance pay may have significantly underestimated the effects in the long run due to missing the sorting component of the effects.

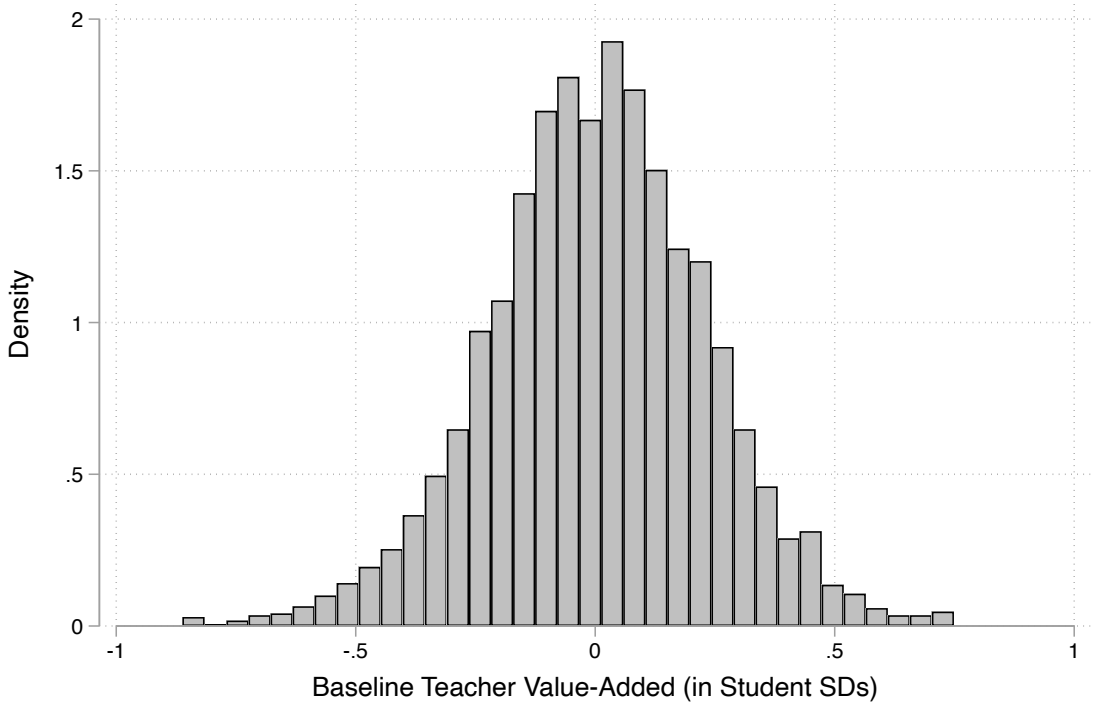
1.11 Figures

Figure 1.1: Experiment Timeline



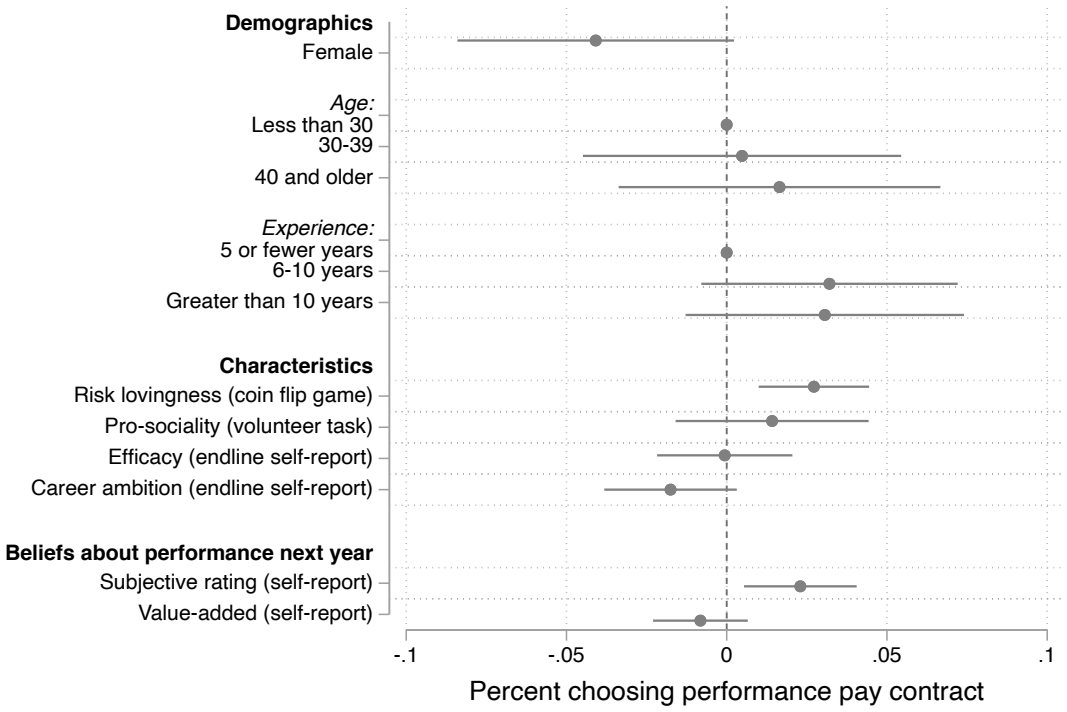
Notes: The figure presents the experimental timeline from June 2017 through May 2019. Our measure of ability comes from the calculation of teacher value-added in June 2017 prior to the introduction of the treatments. Our measure of the behavioral effect of performance pay comes from comparing the treatment and control sample in January 2019, a year after the introduction of the new contracts. We measure teacher's job choices twice: first, from the contract choice elicitation exercise, and second, from where they choose to work starting in August 2018, a semester after the treatments have been announced.

Figure 1.2: Distribution of Teacher Value-Added at Baseline



Notes: This figure presents the distribution of teacher value-added for 3,687 teachers in the school system at baseline. Teacher value-added is calculated using administrative test score data from June 2016 and June 2017 (the two years prior to the intervention). Estimates are calculated following Kane and Staiger [2008], using an empirical Bayes approach.

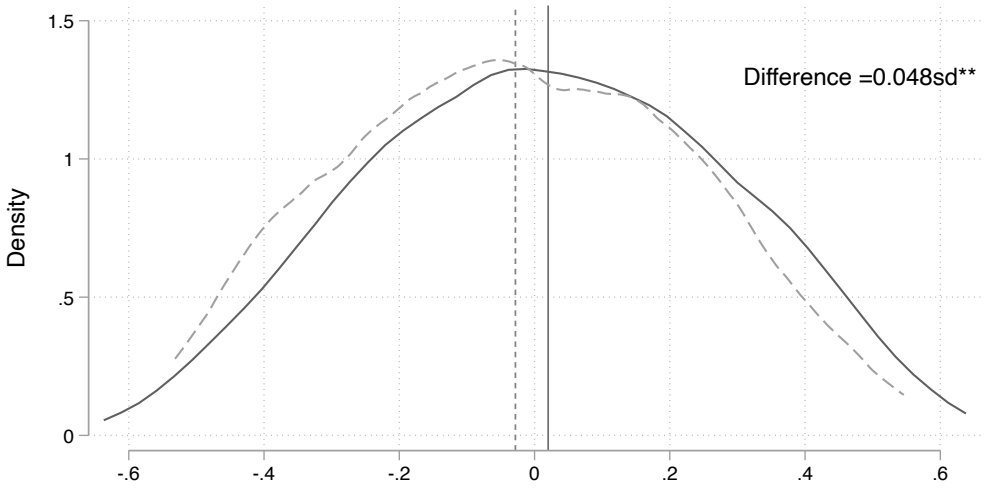
Figure 1.3: Predictors of contract choice



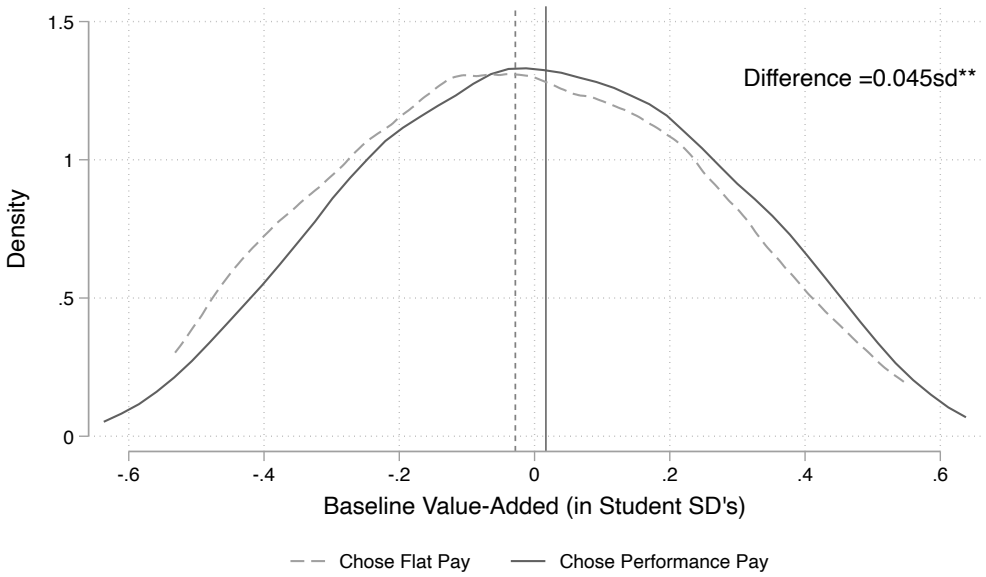
Notes: This figure presents coefficients and 95% confidence intervals of bivariate regressions of teacher's contract choice on teacher demographics, characteristics and beliefs. Teacher's contract choice is a dummy for whether they selected a performance pay or flat pay contract. All independent variables, other than gender, age and experience, are standardized z-scores. Data is at the teacher-decision level, as teachers are asked to choose between performance and flat pay, first using an objective performance measure, then a subjective performance measure. Demographic data come from school administrative records. Characteristics (except efficacy and career ambition), beliefs and contract choice come from a baseline survey with 2,481 teachers.

Figure 1.4: Distribution of Baseline Value-Added by Contract Choice

Panel A: Objective Performance Metric



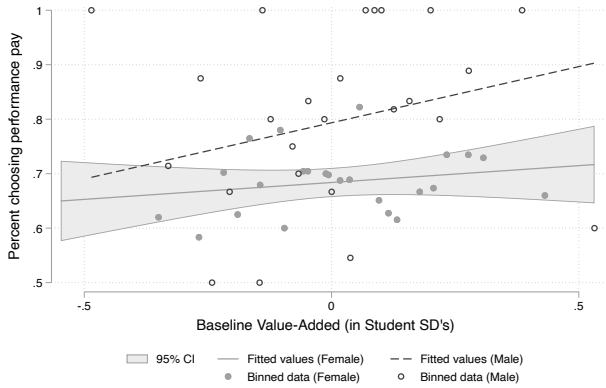
Panel B: Subjective Performance Metric



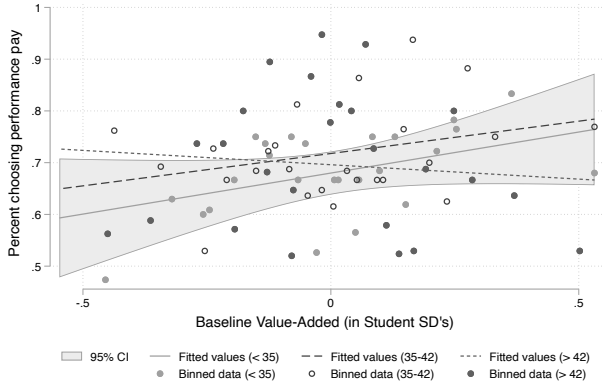
Notes: This figure plots the distribution of baseline teacher value-added for teachers who chose performance pay (solid line) versus flat pay (dotted line). Panel A presents results for the choice between objective (value-added based) performance pay versus flat pay. Panel B presents results for the choice between subjective (principal evaluation based) performance pay versus flat pay. Choice data comes from the contract choice exercise conducted in October 2017. Value-added is calculated using two years of administrative data prior to the start of the intervention. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Figure 1.5: Relationship between Value-Added and Contract Choice by Demographics

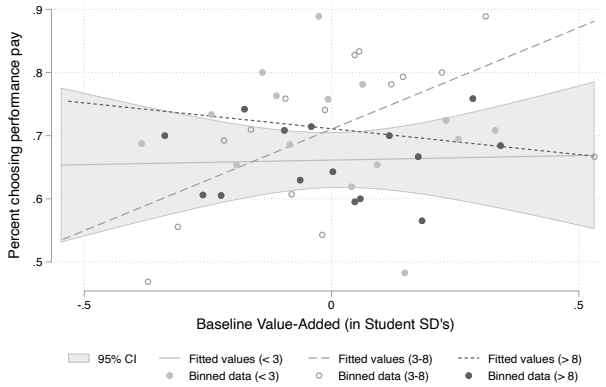
Panel A: By Teacher Gender



Panel B: By Teacher Age



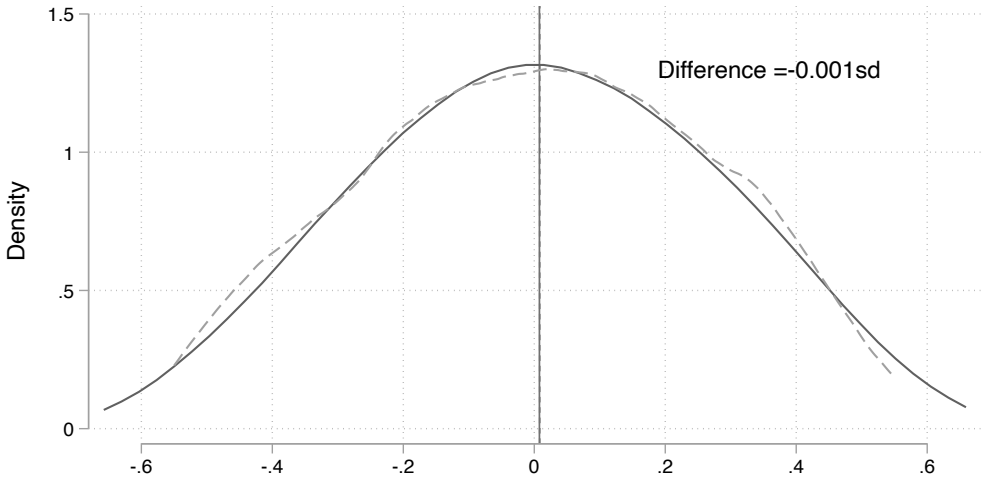
Panel C: By Teacher Experience (years)



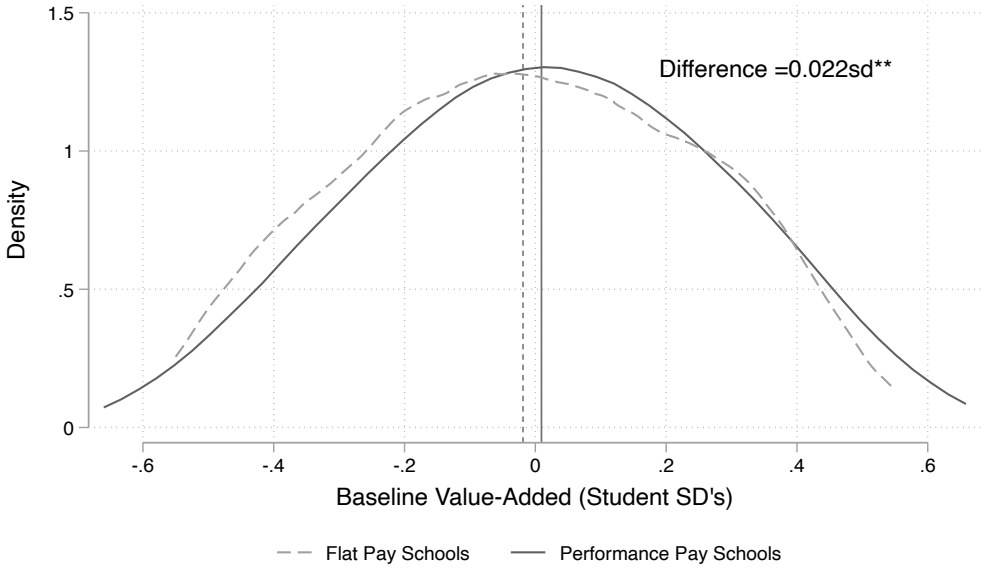
Notes: These figures plot the relationship between teacher quality as measured by baseline value-added and teacher's contract choice. The graph plots binned values of *Teacher Baseline Value-Added* by the percent of teachers in that bin that chose performance pay. Results are shown by teacher characteristic. Choice data comes from the contract choice exercise conducted in October 2017. Value-added is calculated using two years of administrative data prior to the start of the intervention.

Figure 1.6: Distribution of Teacher Baseline Value-Added by School and Year

Panel A: December 2017 (Baseline)

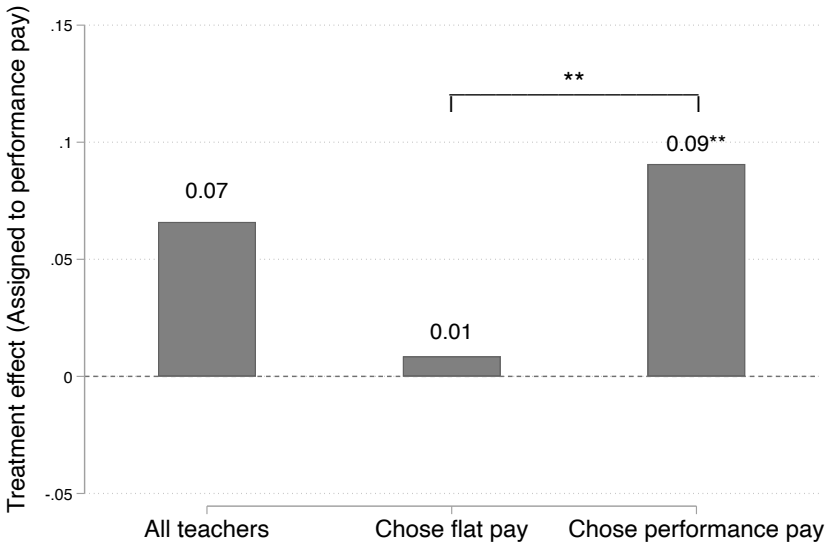


Panel B: December 2018 (One year after treatment announcement)



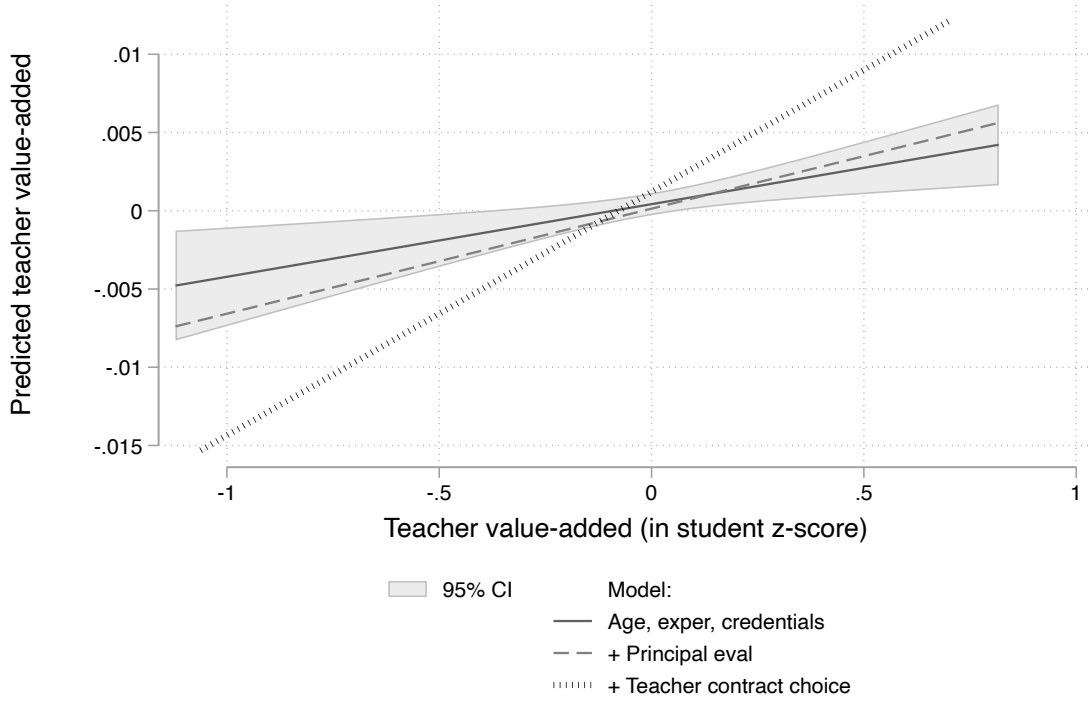
Notes: These figures plots the distribution of baseline teacher value-added for teachers in performance pay versus flat pay schools. Panel A provides the distribution in December 2017 (one month before the treatments are announced). Panel B provides the distribution in December 2018 (11 months after the treatments are announced). Teacher employment data comes from school administrative records. Value-added is calculated using two years of administrative data prior to the start of the intervention. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Figure 1.7: Treatment Effect by Contract Choice



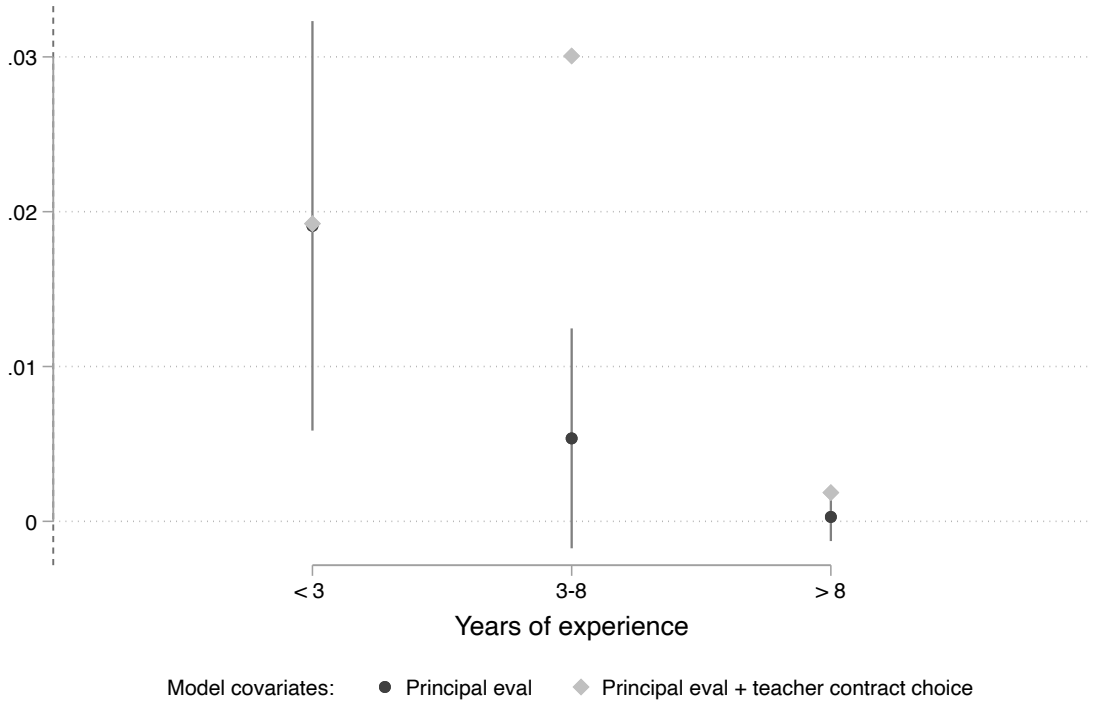
Notes: This figure presents the treatment effects from the performance pay on endline test scores. The first bar presents the effects for all teachers. The second bar presents the treatment effects for teachers who stated in the baseline contract choice exercise that they wanted a flat pay contract. The third bar presents the effects for teachers how wanted a performance pay contract. Endline test scores come from a test conducted by the research team with students in class 4-13 in five subjects in January 2019. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Figure 1.8: Predicting Teacher Value-Added



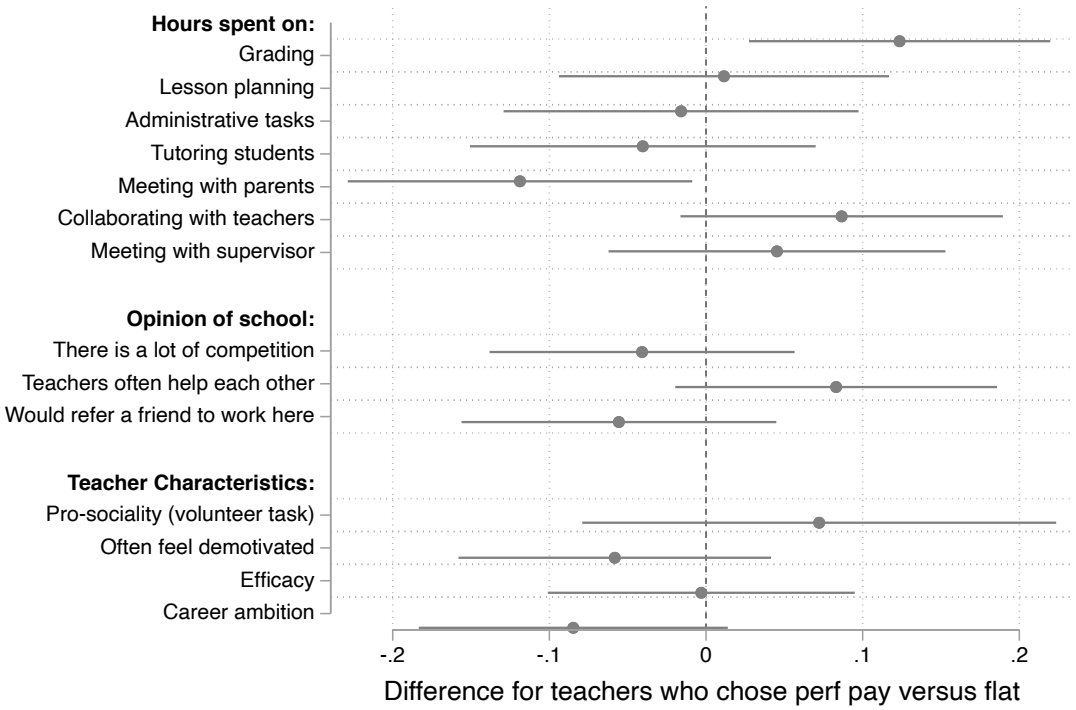
Notes: This figure presents the relationship between value-added and predicted value-added for three different models. The first model (solid line) just includes teacher demographics (age, experience and credential-type fixed effects). The second model (dashed line) uses demographics and principal evaluation. The third model includes demographics, principal evaluation and teacher's baseline contract choice.

Figure 1.9: Predicting Teacher Value-Added by Experience



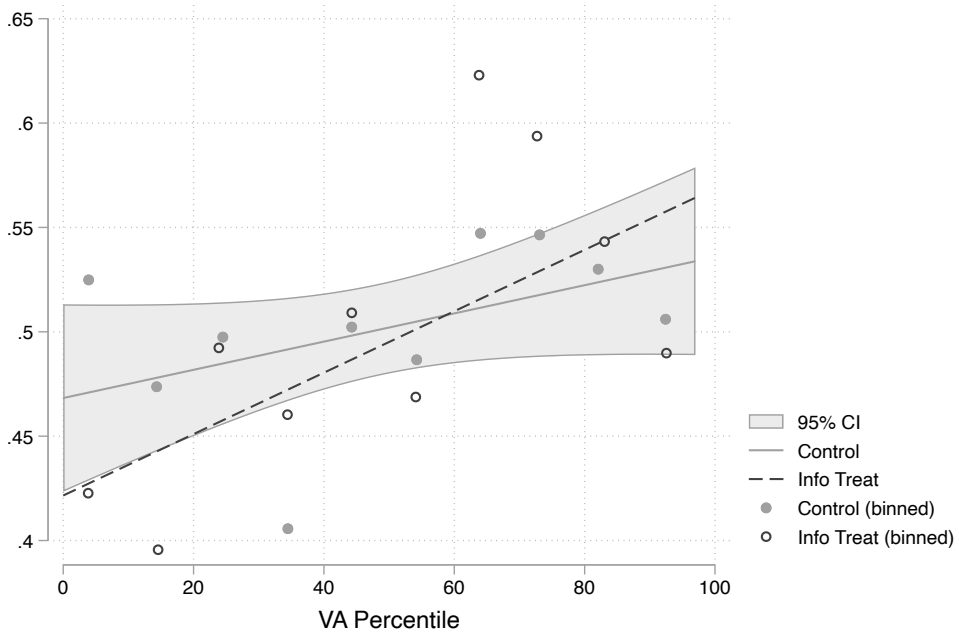
Notes: This figure presents the coefficient and 95% confidence intervals for predicted value-added on value-added for two different models. The first model (black circle) uses principal evaluation. The second (gray diamond) model includes principal evaluation and teacher's baseline contract choice. Results are presented by teacher experience level.

Figure 1.10: Predictors of contract choice



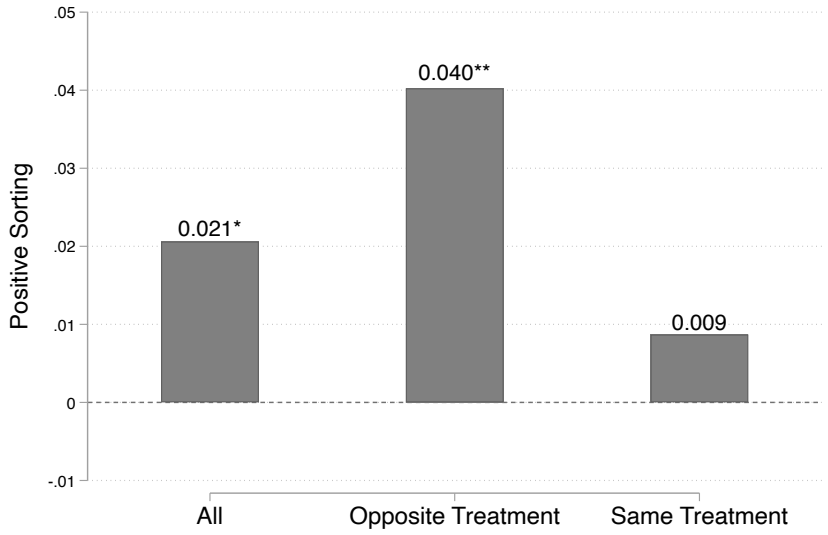
Notes: This figure presents coefficients and 95% confidence intervals of bivariate regressions of teacher time use and characteristics on teacher's contract choice on. Teacher's contract choice is a dummy for whether they selected a performance pay or flat pay contract. All outcomes are standardized z-scores. Data is at the teacher-decision level. Teachers are asked to choose between performance and flat pay, first using an objective performance measure, then a subjective performance measure. Teacher time use and characteristics come from the endline teacher survey.

Figure 1.11: Beliefs and Contract Choice by Teacher Value-Added



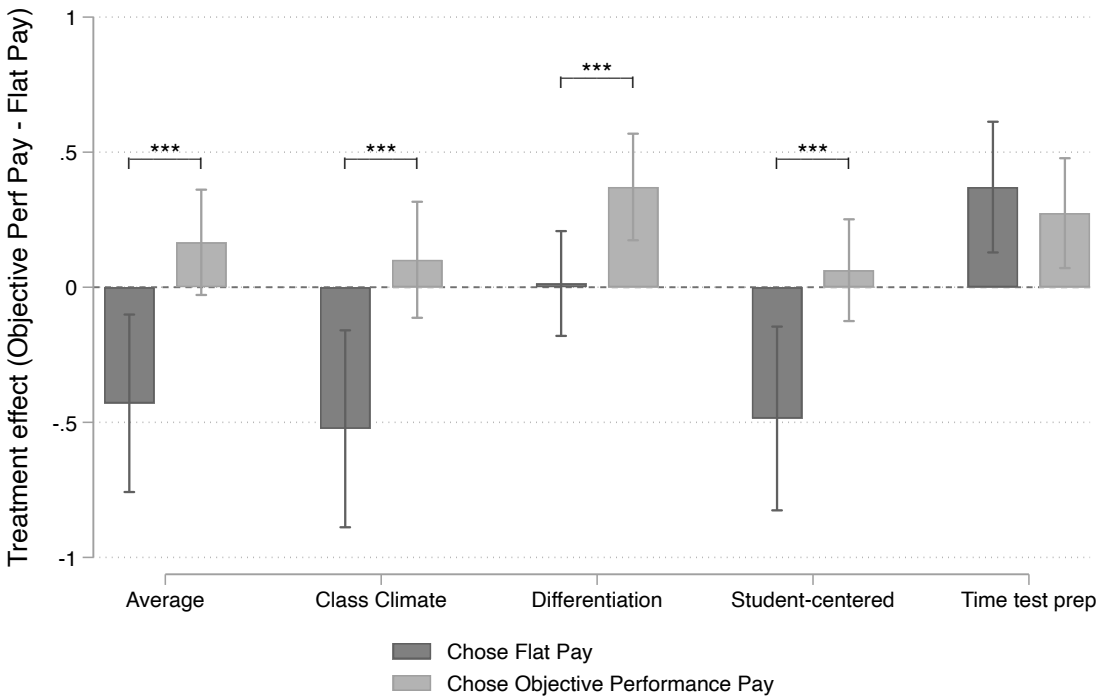
Notes: The figure shows the relationship between teacher's contract choice and their value-added. The solid line, 95% confidence interval, and circles present the relationships for control teachers. The dotted line and white circles show the relationship for teachers who received information about their value-added in the previous year. Belief and choice data come from the baseline survey conducted in October 2017. Value-added is calculated using two years of administrative data prior to the start of the intervention. The information treatment was conducted during the baseline survey.

Figure 1.12: Positive Sorting by Closest School's Treatment



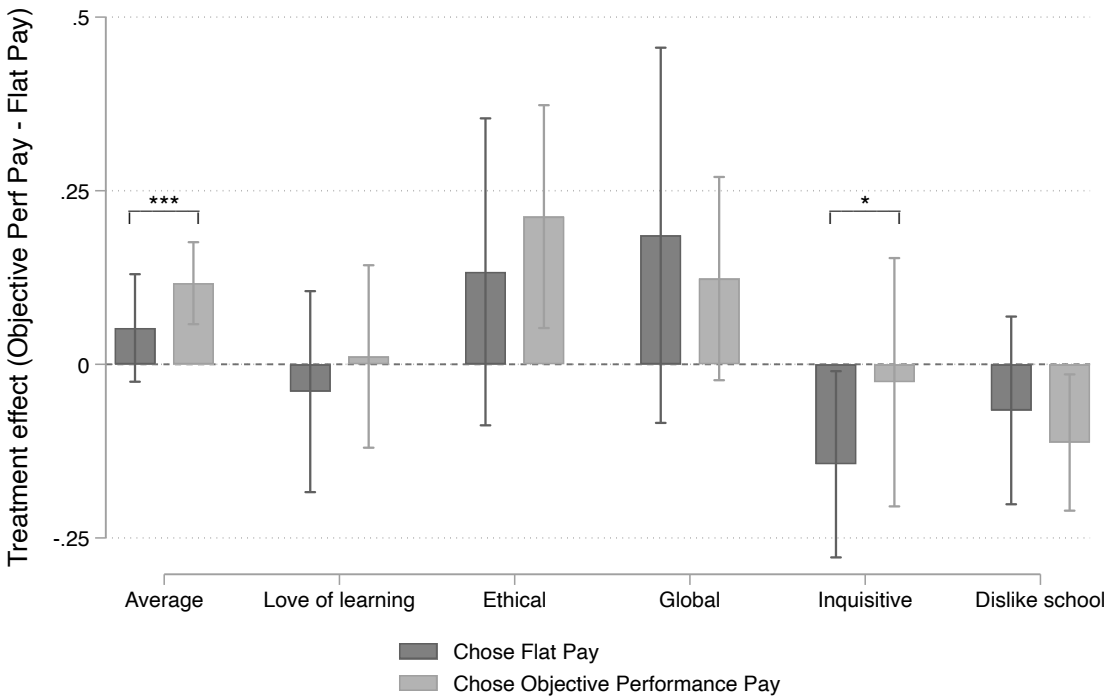
Notes: This figure presents the difference in baseline value-added among teachers employed at performance pay versus flat pay schools at endline. The first bar presents the results for all teachers. The second presents the results for teachers whose closest school to them was assigned the opposite treatment as they were assigned. The last bar presents results for teachers whose closest school received the same treatment as the teacher was assigned. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Figure 1.13: Treatment Effects on Classroom Observations by Contract Choice



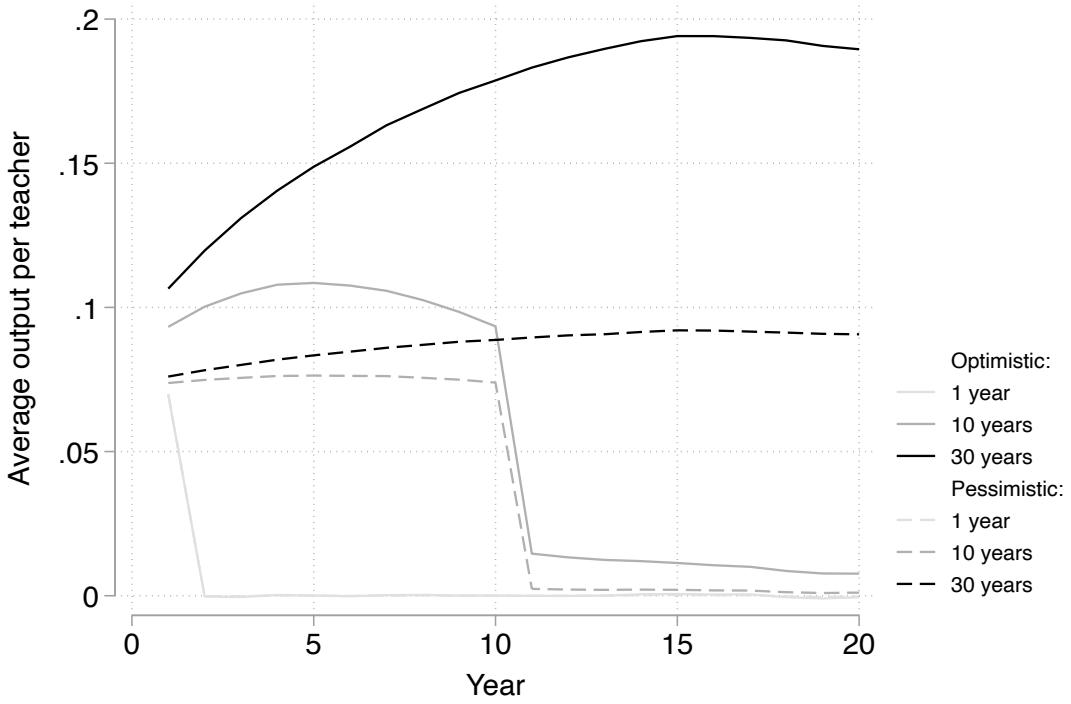
Notes: This figure presents the treatment effect and 95% confidence intervals of objective performance pay relative to flat pay for teachers who chose flat pay (left bar) versus chose performance pay (right bar). Outcomes are from classroom observation data. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Figure 1.14: Treatment Effects on Student Surveys by Contract Choice



Notes: This figure presents the treatment effect and 95% confidence intervals of objective performance pay relative to flat pay for teachers who chose flat pay (left bar) versus chose performance pay (right bar). Outcomes are from student endline survey data. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Figure 1.15: Policy Simulations



Notes: This figure presents the results of the policy counterfactual simulations. It shows the effect of introducing a 1 year, 10 year or 30 year performance pay policy on the average output per teacher. The solid lines use the optimistic parameter values and the dashed lines use the pessimistic parameter values.

1.12 Tables

Table 1.1: Descriptive Statistics about Study Sample and Comparison Sample

	Study Sample		Private Schools		Public Schools	
	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Teacher Characteristics						
Age	35.1	9.0	25.3	7.5	39.9	9.0
Female	0.81	0.40	0.78	0.42	0.45	0.50
Years of experience	9.9	6.7	4.8	7.1	16.2	10.4
Has BA	0.95	0.22	0.33	0.47	0.55	0.50
Salary, USD (PPP)	13,000	5,000	1,400	1,100	7,800	3,600
Panel B. Principal and School Characteristics						
Female	0.72	0.42	0.49	0.50	0.30	0.46
Overall management score	4.27	0.43	1.78	0.34	1.61	0.34
People management score (out of 5)	4.14	0.53	1.83	0.35	1.70	0.38
Operations management score (out of 5)	4.32	0.61	1.71	0.42	1.40	0.38
Students per school	841	581	1320	997	967	756
Student-teacher ratio	31.8	12.4	27.5	12.8	33.6	24.7

Notes: This table reports summary statistics on teacher, principal and school characteristics for our study sample, and a comparison sample in Pakistan (Panel A) and India (Panel B). Data in panel A, columns (1) and (2) comes from administrative data provided by our partner school system. Data in panel B, columns (1) and (2) is from an endline survey conducted with 189 principals and vice principals and 5,698 teachers in our study sample. Data in panel A, columns (3)-(6) comes Learning and Educational Achievement in Pakistan Schools (LEAPS) data set [Bau and Das, 2020]. Data in panel B, columns (3)-(6) is from the World Management Survey data conducted by the Centre for Economic Performance [Bloom et al., 2015]. We restrict to the 318 schools located in India from that sample.

Table 1.2: Teacher Value-Added by Contract Choice

	Teacher Baseline Value-Added (in Student SDs)			
	(1)	(2)	(3)	(4)
Chose Performance Pay	0.0485**	0.0450**	0.0452**	0.0387*
	(0.0207)	(0.0207)	(0.0218)	(0.0221)
Principal Rating of Teacher		0.0210**		0.0202*
		(0.0104)		(0.0105)
Observations	1284	1284	1284	1284
Performance Metric	Objective	Objective	Subjective	Subjective
Control Mean	-0.0283	-0.0283	-0.0284	-0.0284
Control SD	0.349	0.349	0.345	0.345

Notes: This table presents the relationship between teacher characteristics and baseline value-added. *Teacher Baseline Value-Added* is measure of teacher value-added using test score data from the two years prior to the intervention. It is in student standard deviations. *Chose Performance Pay* is a dummy variable for whether a teacher chose performance pay or flat pay during the baseline choice exercise. Columns (1) and (2) present results for the choice between objective (value-added based) performance pay and flat pay. Columns (3) and (4) present results for the choice between subjective (principal evaluation based) performance pay and flat pay. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table 1.3: Teacher Value-Added by Contract Choice and Demographics

	Chose Performance Pay		
	(1)	(2)	(3)
Teacher Baseline Value-Added (in Student SDs)	0.133***	0.215***	0.0777
	(0.0439)	(0.0519)	(0.0790)
Male	0.0804***		
	(0.0240)		
Value-Added * Male	-0.0544		
	(0.113)		
$>$ 40 years old		-0.0114	
		(0.0180)	
Value-Added * $>$ 40 years old		-0.218***	
		(0.0844)	
$<$ 5 years experience			-0.0915***
			(0.0328)
6-10 years experience			-0.0531**
			(0.0264)
Value-Added * $<$ 5 years experience			-0.136
			(0.145)
Value-Added * 6-10 years experience			0.257**
			(0.116)
Constant	0.710***	0.723***	0.731***
	(0.00964)	(0.0117)	(0.0154)

Notes: This table presents the relationship between teacher contract choice and baseline value-added. *Teacher Baseline Value-Added* is measure of teacher value-added using test score data from the two years prior to the intervention. It is in student standard deviations. *Chose Performance Pay* is a dummy variable for whether a teacher chose performance pay or flat pay during the baseline choice exercise. Results are show interacted with teacher characteristics (gender, age, and years of experience). Teacher characteristics come from school administrative data. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table 1.4: Teacher Quality by School

	Teacher Baseline Value-Added (in Student SDs)			
	(1)	(2)	(3)	(4)
Performance Pay Schools	-0.0160	-0.0143	0.00178	0.00347
	(0.0188)	(0.0189)	(0.0198)	(0.0202)
Post	-0.0191*	-0.0194*	-0.0195*	-0.0203*
	(0.0107)	(0.0108)	(0.0108)	(0.0106)
Performance Pay Schools*Post	0.0222**	0.0225**	0.0231**	0.0216*
	(0.0113)	(0.0113)	(0.0113)	(0.0112)
Principal Rating of Teacher				0.0201***
				(0.00711)
Randomization Strata FE	Yes	Yes	Yes	Yes
Grade and Subject FE		Yes	Yes	Yes
Region FE			Yes	Yes
Control Mean	0.0190	0.0190	0.0190	0.0187
Control SD	0.327	0.327	0.327	0.329
Clusters	243	243	243	239
Observations	6991	6991	6991	6747

Notes: This table presents the relationship between teacher quality (as measured by teacher value-added) and where teachers choose to work. The outcome is *Teacher Baseline Value-Added*, measured using test score data from the two years prior to the intervention. *Performance Pay School* is a dummy for if a teacher works at a school that is assigned to a performance pay treatment contract (as compared to works at a school which was assigned a control flat pay contract). *Post* is a dummy that is equal to 0 in December 2017 and 1 in December 2018. Data is at the teacher-year level. Column (1) presents basic specification (eq. 1.10). Columns (2)-(4) add additional controls. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table 1.5: Treatment Effect by Contract Choice

	Endline Test (z-score)					
	(1)	(2)	(3)	(4)	(5)	(6)
Assigned Perf Pay Treat	0.0881**	0.0660	0.00857	0.00837	0.0630	0.00160
	(0.0397)	(0.0408)	(0.0511)	(0.0511)	(0.0421)	(0.0551)
Chose Perf Pay* Assigned Perf Pay Treat			0.0822**	0.0824**		0.0882**
			(0.0406)	(0.0405)		(0.0440)
Principal Rating of Teacher				0.00323		
				(0.00989)		
Baseline Value-Added*Assigned Perf Pay Treat					-0.0729	-0.0854
					(0.129)	(0.129)
Control Mean	-0.00377	7.94e-10	7.94e-10	7.94e-10	-0.00223	-0.00223
Control SD	0.999	1.000	1.000	1.000	0.997	0.997
Clusters	190	114	114	114	109	109
Observations	494956	144009	144009	144009	126989	126989
Randomization Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Baseline	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table presents the treatment effect of performance pay contracts on endline test scores by teacher characteristics. The outcome is students' standardized z-score from the endline test conducted in January 2019 at the exam-student-teacher level. *Assigned Perf Pay Treat* is a dummy for whether a teacher taught at a school assigned to performance pay at baseline. *Chose Perf Pay* is a dummy variable for whether a teacher chose objective performance pay or flat pay during the baseline choice exercise. *Principal Rating of Teacher* is the baseline subjective rating z-score of the teacher by their principal. Column (1) presents the treatment effect for all teachers. Column (2) presents treatment effects for the 30% of teachers who were part baseline survey and choice exercise. Column (3) and (5) presents heterogeneity in treatment effect by contract choice and value-added, respectively. Column (6) combines the two and column (4) controls for principal's beliefs about teacher quality. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table 1.6: Principal Beliefs about Teacher Quality

	Principal Belief (z-score)									
	(1) All	(2) Attendance	(3) Discipline	(4) Analysis	(5) VA	(6) All	(7) All	(8) All	(9) All	(10) All
Teacher Outcome (z-score)	0.168***	0.192***	0.231**	0.136	-0.0435	0.238***	0.0580	0.184***	0.173***	0.150***
	(0.0433)	(0.0503)	(0.104)	(0.125)	(0.0831)	(0.0661)	(0.0680)	(0.0482)	(0.0498)	(0.0383)
Principal experience (years)						0.0160***			0.0159***	
						(0.00516)			(0.00542)	
Teacher Outcome*Principal experience						-0.00656				
						(0.00496)				
Observation treatment							-0.0433			
							(0.0900)			
Teacher Outcome*Observation treatment							0.195*			
							(0.1000)			
Overlap $>$ 2 years with teacher								0.164*	0.0887	0.110
								(0.0851)	(0.0887)	(0.0977)
Teacher Outcome*Overlap $>$ 2 years								-0.175**	-0.161*	-0.150**
								(0.0804)	(0.0828)	(0.0703)
Observations	702	250	143	143	166	702	594	702	698	702
Grade Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Principal Fixed Effects	No	No	No	No	No	No	No	No	No	Yes

Notes: This table presents the relationship between teacher outcomes and principals beliefs about those outcomes. There are four outcomes principals rate teachers on: attendance, management of student discipline, incorporation of analysis and inquiry skills and value-added. *Principal beliefs* are from principal endline survey data. Actual teacher outcomes come from administrative and classroom observation data. Attendance is measured using biometric clock in and out data. Discipline and analysis/inquiry are rates via classroom observations. Column (2)-(5) separates the results by outcome type. Columns (6)-(10) add interactions with principal characteristics. *Principal experience* is the number of years the principal has worked in the school system. *Observation treatment* is a dummy for whether the teacher was assigned to be observed more frequently by their principal. This treatment was in place from September 2018 to January 2019. *Overlap $>$ 2 years* is a dummy for whether the teacher and principal have worked together at the same school for at least two years. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table 1.7: Teacher Value-Added by Contract Choice - Information Treatment

	Percentile Rank
Choose Perf Pay	6.807***
	(0.777)
Info Treatment	-1.959*
	(1.138)
Choose Perf Pay*Info Treatment	2.953*
	(1.582)
Control Mean	45.93
Control SD	27.08
Observations	6916

Notes: This table presents the relationship between teacher contract choice and baseline value-added for those that received the information treatment. *Percentile Rank* is teacher's percentile rank within their school. *Choose Performance Pay* is a dummy variable for whether a teacher chose performance pay or flat pay during the choice exercise. *Info Treatment* is a dummy for whether the teacher received information about their performance in the previous year. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table 1.8: Positive Sorting by Closest School's Treatment

	Teacher Baseline Value-Added (in Student SDs)			
	(1)	(2)	(3)	(4)
Performance Pay Schools	-0.0121	-0.0546	-0.00275	-0.0213
	(0.0182)	(0.0447)	(0.0396)	(0.0374)
Post	-0.0185*	-0.0243*	-0.00551	0.00188
	(0.0108)	(0.0139)	(0.0263)	(0.0257)
Perf Pay Schools*Post	0.0206*	0.0403**	0.00870	0.00119
	(0.0114)	(0.0183)	(0.0270)	(0.0263)
Opposite Treat				0.00973
				(0.0443)
Perf Pay Schools*Opposite Treat				-0.0233
				(0.0510)
Post*Opposite Treatment				-0.0265
				(0.0273)
Post*Perf Pay Schools*Opposite Treat				0.0392
				(0.0299)
Sample	All	Opposite	Same	
Randomization Strata FE	Yes	Yes	Yes	Yes
Control Mean	0.0190	0.0190	0.0190	0.0190
Control SD	0.327	0.327	0.327	0.327
Clusters	243	115	172	203
Observations	6991	1211	3495	4706

Notes: This table presents the extent of positive sorting for teachers who faced different switching costs. The outcome is *Teacher Baseline Value-Added*, measured using test score data from the two years prior to the intervention. *Performance Pay School* is a dummy for if a teacher works at a school that is assigned to a performance pay treatment contract (as compared to works at a school which was assigned a control flat pay contract). *Post* is a dummy that is equal to 0 in December 2017 and 1 in December 2018. Data is at the teacher-year level. Column (1) presents the results for all teachers. Column (2) presents the results for teachers whose closest neighboring school was assigned the opposite treatment as their school (low switching cost). Columns (3) presents the results for teachers whose closest neighboring school had the same treatment as them (high switching costs). Standard errors are clustered at the school level. $^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$.

Table 1.9: Values of Key Parameters

Parameter	Value		Source/Calculation
	Pessimistic	Optimistic	
Mean ability, μ_θ	0	-	Test scores
SD ability, σ_θ	0.15	0.30	Test scores
Mean behavioral effect, μ_β	0.07	-	Test scores
SD behavioral effect, σ_β	0.14	0.28	Test scores
Covariance θ and β, $\rho_{\theta\beta}$	$-1.94\text{x}10^{-4}$	-	Test score
Fraction new entrants	0.3	-	Admin data
Job-employee specific utility, σ_ϵ	\$360	-	Admin data
Job-employee time shocks, σ_e	\$180	-	Admin data
Mean cost to change professions, μ_c	\$1,120	-	Survey
SD cost to change professions, σ_c	\$1,200	-	Survey
Accuracy of priors (existing teachers) θ, α_θ	0.067	0.100	Admin data/survey
Accuracy of priors (existing teachers) β, α_β	0.035	0.053	Admin data/survey
Accuracy of priors (non-teachers) θ, α_θ	0.046	0.070	Admin data/survey
Accuracy of priors (non-teachers) β, α_β	0.024	0.046	Admin data/survey
Length of time on job (exponential function), τ_i	10		Admin data

Notes: This table reports the parameter values used in the policy counterfactual simulations.

Chapter 2

Subjective versus Objective Incentives and Teacher Productivity

2.1 Introduction

How should schools incentivize teachers when effort is non-verifiable or non-contractable? Contract theory provides an answer. The second best is to incentivize on outcomes of the employee's production function. However, this introduces two new problems – distortion, over-incentivizing measurable outcomes while ignoring others, and noise, outcomes are a noisy function of employee effort. How do most non-schools actually incentivize workers? They use manager-discretionary (subjective) incentives rather than outcome-based (objective) ones. Raises, promotions, and terminations are subject to manager discretion for most employees. In the US, 85% of full-time employees have at least one aspect of their compensation determined by their manager, and 90% of teacher performance evaluations have a subjective component [Engellandt and Riphahn, 2011, National Center for Education Statistics, 2011]. Despite the prevalence of subjective incentives, there is limited causal evidence on the effect of these incentives and whether they could work in the teaching setting.

In this paper, we ask two questions: What is the effect of subjective versus objective incentives on teacher productivity? Are subjective incentives able to help alleviate problems of noise and distortion, which often plague objective incentives? We answer these questions by conducting an 18-month randomized controlled trial with 234 private schools in Pakistan. We randomize schools to provide core teachers with one of three contracts: (i). control: flat raise – all teachers receive a raise of 5% irrespective of performance, (ii). treatment 1: subjective performance raise – teachers receive a raise from 0-10% based on their manager's rating of their performance,[1] or (iii). treatment 2: objective performance raise – teachers

[1]Managers are generally principals or vice-principals and spend about a third of their time on employee

receive a raise from 0-10% based on their students' mid-year and end of year test performance [Barlevy and Neal, 2012]. Both treatments are within-school tournaments and have the same distribution of raise thresholds. These similarities allow us to isolate the effort response from just changing the performance metric (manager rating versus test score) while holding other features of the incentive structure constant.

We use detailed administrative, survey, test, and classroom observation data to understand each contract's effect on teacher effort and student outcomes. Student outcomes are measured along two dimensions: test scores and socio-emotional development. Test score data comes from an endline test conducted by the research team, one month after the end of the contract. Students are tested in core subjects (English, Urdu, math, science, and economics) in grades 4-13. A variety of question types and sources allow us to test whether effects are driven by memorization-type questions. Socio-emotional development is measured along four dimensions: love of learning, ethical behavior, inquisitiveness, and global competency. These dimensions are measured using self-report survey items drawn from several psychological indices used for measuring socio-emotional development in children.[2]

In our first main result, we show that both subjective and objective contracts are equally effective at increasing test scores. Both contracts increase test scores by 0.09 sd, which is very similar to average effects from meta-analyzes of performance pay for teachers [Pham et al., 2020]. These results are consistent across subject and grade and are not driven by rote-memorization type questions. However, we find, in contrast to the test score results, objective and subjective incentives have different effects. Objective incentives negatively affect student socio-emotional development, including a significant decrease in love of learning and an increased likelihood students say they want to change schools. Subjective incentives result in a small positive effect on overall socio-emotional skills. These combined effects suggest that teachers under objective contracts focused exclusively on improving student academic improvement, at the cost of more well-rounded development for students. Whereas, teachers under the subjective contract were able to prioritize both areas.

To understand teachers' behavioral responses to these incentive contracts, we compile rich data on teacher behavior inside and outside the classroom. We record 6,800 hours of classroom footage and review it using a standard classroom observation rubric [Pianta et al., 2012]. The rubric captures teacher behavior along dozens of dimensions, from the use of punitive discipline to the proportion of student versus teacher talk time. The rubric also measures the amount of time spent on test-taking or test-preparation activities. To measure effort outside the classroom, we have teachers complete a time use questionnaire. Combined these two data sources allow us to understand teacher behavior change under subjective versus objective incentives.

In our second main result, we find both subjective and objective incentives lead to changes in classroom practices. As one might expect, subjective incentives spur actions that managers

management tasks, such as observations, feedback, and professional development.

[2]Items are drawn from the National Student Survey, Learning and Study Strategies Inventory, Big Five (children's scale), Eisenberg's Child-Report Sympathy Scale, Bryant's Index of Empathy Measurement, Afrobarometer, World Values Survey, and Epistemic Curiosity Questionnaire.

value, and objective incentives spur actions that most quickly and easily translate into test score gains. Subjective incentives lead to increased targeting of individual student needs within the classroom and the use of technology in the classroom. Both teaching practices are one's principals identified as markers of high-quality teaching. Objective incentive schools see a five-fold increase in class time on test preparation activities. These teachers also exhibit more negative discipline techniques, such as yelling at students.

Our reduced form effects suggested that subjective performance incentives increase teacher effort without producing distortionary effects. How are managers able to accomplish this? We find on average managers place significant value on teachers value-added and pedagogy. We also do not find any evidence of favoritism or gender bias. However, there is heterogeneity in managers' application of the contract. We cannot reject there is no effect of subjective performance pay for the worst quintile of managers.

We then draw on the model of moral hazard with multi-tasking to explain our reduced form results: i). similar, positive effects of subjective and objective incentives on test scores, ii). negative effects of objective incentives on socio-emotional development, iii). significant differences in teacher classroom behavior across the two treatments. Moral hazard models with multi-tasking [Baker, 2002] isolate two main components of the incentive structure which affect employee response: noise (correlation between employee action and incentive pay) and distortion (correlation between piece rate for different actions and marginal return to those actions on firm outcomes). Our paper seeks to understand whether noise and distortion serve as important mechanisms of the reduced form effects we see.

Our empirical approach for this mechanism analysis proceeds in three steps. First, we show differences in employee's perception of the noisiness and distortion for subjective versus objective incentives. Second, we exploit partially exogenous heterogeneity within a given treatment to isolate the causal effect of noise and distortion each individually on student outcomes. Finally, we bring those two estimates together and show that given the difference in levels of noise and distortion across the contracts and the effect of noise and distortion on student outcomes, we can explain a large portion of the reduced form effects through these channels. We explain each step in detail below.

The first step is showing that teachers believe there are differences in the extent of noise and distortion across the two treatments. We do this by asking teachers at endline the extent to working harder will increase their incentive pay. If they believe their effort closely maps into their pay then this is a *less* noisy incentive system. Then we ask what types of actions (lesson planning, improving pedagogy, helping other teachers, etc) are rewarded under each system. This allows us to measure teachers perception of whether the incentive is distorted toward certain student outcomes at the cost of others.

We find that teachers believe subjective performance incentives are *less* noisy than objective incentives, and, therefore, view subjective incentives as more effective at motivating behavior. They view test-score based incentives as much less within their control because so many other factors beyond their effort affect student scores. We also find that teachers in the objective treatment are more likely to prioritize the type of actions which lead to test score gains, at the cost of other areas of student development. Teachers under subjective contract prioritize actions that lead to academic gains and also prioritize administrative tasks, which

are likely to be preferred by their manager.

We also show there are no other differences beyond noise and distortion across the two treatment arms. We show there is similarity in implementation timelines, understanding of the contract treatments, and beliefs about the fairness of each treatment arm.

The second step of our mechanisms analysis is to demonstrate that noise and distortion themselves affect student outcomes. To do this, we zoom in to the subjective treatment schools and look at settings with high and low noise and then high and low distortion. By controlling for other differences across settings, we are able to isolate the effect of these two mechanisms on outcomes.

To determine the effect of noise on student outcomes, we compare subjective treatment schools with managers whom teachers rate as accurate in assessing teacher effort versus managers rated as inaccurate in assessing teacher effort. We use this rating of managers' accuracy interacted with treatment status as an instrument for the perceived noisiness of the contract. We show that this rating of managers only affects teacher's rating of noisiness in the subjective arm. This instrument for noise is robust to controlling for many other features of the contract and school environments.

Using this instrument for noise, we find that a 1 SD increase in the perceived noisiness of the contract decreases hours worked by 13 and decreases student test scores by 0.2 SD. These results are robust to a variety of controls. This suggests that employees are very sensitive to the noisiness of the contract, and that this affects the success performance pay has in inducing an effort response from employees.

To understand the effect of distortion on student outcomes, we again exploit variation within the subjective performance pay schools. We use data on managers' preferences prior to the start of the experiment. Before the treatments are announced managers sit down with the teachers and delineate goals for the following year for that teacher. Example goals include increasing students' English proficiency, reaching certain grade targets, or improving lesson plans. We code these goals using text analysis and categorize them into four types of teacher actions: administrative tasks, professional development and collaboration tasks, improvements in teacher pedagogy, and test-score based goals. A month after these goals are set between managers and teachers, we announce the treatment assignment.

Of course, schools in which managers focus on administrative goals versus those in which managers focus on pedagogy goals are likely different in many ways. Therefore, our approach is to interact these goals with the subjective treatment, to isolate the effect of these goals in settings in which teachers would be more likely to focus on them (those who were assigned subjective treatment) relative to places where the goals have no financial stake (objective and flat treatment schools). We use the interaction of subjective treatment and goal, controlling for level differences, to isolate the effect of these goal differences on student outcomes. We find that a larger focus on test scores and professional development increases students' endline test scores. However, more focus on test scores results in negative effects on student socio-emotional development. These results are robust to controlling for other features of the contract environment.

Combined, these results help us understand why it is possible to have the same effect on test scores without needing to incentivize test scores directly. Subjective incentives are less

noisy, producing a larger overall response, and less distorted, allowing teachers to prioritize multiple areas of student development. We find that the noise and distortion channel are able to explain a substantial portion of the reduced form effects we see.

Our paper makes three key contributions. First, it is the first study, to our knowledge, to isolate the causal effect of subjective versus objective incentives and the effect of subjective versus flat incentives for employees in any sector [Lazear and Oyer, 2012, Oyer and Schaefer, 2011]. Existing studies have tested bundled incentives (a combined subjective and objective incentives versus no incentives) on employee behavior [Khan et al., 2019, Fryer, 2013]. Previous work has also compared the effect of heterogeneity across plants to measure the effect of more or less steep subjective incentives on employee overtime [Engellandt and Riphahn, 2011]. There is also evidence that managers, especially in educational settings, may have imperfect information about worker effort or may be biased toward certain groups [Jacob and Lefgren, 2008, Gibbs et al., 2004].

Second, we add to a robust literature on the effect of performance pay for teachers by providing two new findings [Lavy, 2007, Muralidharan and Sundararaman, 2011, Fryer, 2013, Goodman and Turner, 2013]. We show the first evidence of objective performance pay having detrimental effects on non-academic student outcomes, consistent with multi-tasking models. Next, we show direct evidence that objective incentives result in teachers distorting their effort toward teaching pedagogy that impacts test performance at the cost of other areas of student development. This includes the use of class time doing test prep and the use of punitive discipline. Both of these results have long been suspected, but we provide the first documentation of such effects [Baker, 2002, Leigh, 2013].

Third, we provide, what we believe is, the first evidence on measuring the extent of noise and distortion within an employee's contract and isolating the effects of those mechanisms on firm outcomes. There is a rich theoretical literature on the importance of these mechanisms [Baker, 2002]. Empirical work has also investigated the role of noise on employee response [Prendergast, 1999, Prendergast and Topel, 1993, Prendergast, 2007].

The remaining sections are organized as follows. Section 2.2 gives an overview of the standard moral hazard model with multi-tasking and highlights the two key mechanisms which underpin the reduced form effects we find. Section 2.3 details the treatment and control conditions, the data collected, and standard implementation checks. Section 2.4 provides the main results of subjective and objective performance incentives on teacher effort and student outcomes. Section 2.5 unpacks the mechanisms underlying the main effects in light of the moral hazard model, and section 2.6 concludes.

2.2 Theoretical Framework

The experimental design is motivated by a model of moral hazard with multi-tasking, as presented in Baker (2002). This theoretical framework helps us rationalize the teacher behaviors and student outcomes we see as a result of each performance incentive. In this section, we lay out this standard model, demonstrate how this translates to the teaching context, and map out how the experimental design connects to the model.

Moral Hazard with Multi-tasking

The firm, a school, produces a single outcome – human capital, $H(\mathbf{a}, e)$ – through a simple linear production function:

$$H(\mathbf{a}, e) = \mathbf{f} \cdot \mathbf{a} + e = f_1 a_1 + f_2 a_2 + ... + e \tag{2.1}$$

Human capital is a function of an n-dimensional vector of actions teachers can take, $\mathbf{a}$, and the n-dimensional vector of marginal products of those actions, $\mathbf{f}$. Human capital is also a function of many other things outside the teacher's action set (environment, parental support, peers, etc.), which are captured by the noise term, e, which is mean zero and has a variance of σ_e^2.

Schools cannot perfectly observe all components of $\mathbf{a}$, but they can observe some features of human capital (for example, test scores) and some actions (for example, teacher attendance). Schools construct a performance contract that pays teachers based on a performance measure, $P(\mathbf{a}, \phi)$, which could be a combination of observable outputs (test scores, student attendance, etc.) and/or actions (teacher attendance, lesson plans, etc.). Teacher's performance measure, and therefore their pay, then is:

$$P(\mathbf{a}, \phi) = \mathbf{g} \cdot \mathbf{a} + \phi = g_1 a_1 + g_2 a_2 + ... + \phi \tag{2.2}$$

The performance measure, $P(\mathbf{a}, \phi)$, is a function of teacher's actions, $\mathbf{a}$, and the marginal return to those actions on the performance measure, $\mathbf{g}$. In effect, $\mathbf{g}$ translates to a piece-rate for each action. ϕ captures everything outside the teacher's actions, which affect the performance measure. It is mean zero and has variance σ_ϕ^2. Two types of noise are captured by ϕ. First is noise coming from features of the performance measure, which are outside the teacher's control. For example, if the performance measure is students' test scores, this could be the students' home environment. Second is the noise coming from mis-measurement of a given action, a_n. For example, if the performance measure is teacher attendance, but principals have error-ridden records of attendance, then this contributes to the noisiness of the performance measure.

Teacher's utility is a function of their pay and a quadratic cost of effort.[3]

$$u(\mathbf{a}, \phi) = \mathbf{g} \cdot \mathbf{a} + \phi - \Sigma_{i=1}^{n} \frac{a_i^2}{2} \tag{2.3}$$

Teachers choose the optimal set of actions that maximizes their utility. Taking the derivative of Eq. 2.3, we have that the optimal decision is to set each action amount equal to the piece rate, $a_1^* = g_1, a_2^* = g_2, ...a_n^* = g_n$.

Given teacher's optimal action set, the average human capital produced by each teacher is:

$$E[H(\mathbf{a}^*, e)] = \mathbf{f} \cdot \mathbf{g} = \|f\|\|g\| cos\theta \tag{2.4}$$

[3]Baker (2002) assumes risk-averse agents with a utility function of $u(\mathbf{a}, \phi) = E[P] - rvar[P] - \Sigma_{i=1}^{n} \frac{a_i^2}{2}$. Because we are not focused on teacher retention, we leave out the risk aversion component, which only enters in determining the nature of the participation constraint and does not affect effort response once an employee has selected the contract.

Average human capital then is a function of the length of the marginal production on human capital vector, $\|f\|$, the length of the piece-rate vector, $\|g\|$, and the alignment between these two vectors, $cos(\theta)$. In other words, human capital is increasing in the steepness of the incentives and how aligned those piece rates are with the human capital production function.

We now go beyond Baker [2002] by making one additional assumption relevant in our context. We can further re-arrange the expression to show the effect that noise in the performance measure has on average human capital. Taking the variance of Eq. 2.2, we have $var(P) = \|g\|^2 var(\mathbf{a}) + \sigma_\phi^2$. Re-arranging, we can substitute this in for $\|g\|$ into Eq. 2.4. Average human capital then is:

$$E[H^*(\mathbf{a}^*, e)] = \|f\| \frac{\sqrt{var(P) - \sigma_\phi^2}}{\sqrt{var(\mathbf{a})}} cos\theta \tag{2.5}$$

Here $\|f\|$ and $var(\vec{a})$ are constant across the two types of performance measures, subjective and objective, we will be comparing. In addition, due to the design of our subjective and objective incentives, $var(P)$, is also constant across the two schemes.[4]

Theoretical Predictions

We are then left with two components of the performance measure that affect average human capital. The key predictions of the model are that average human capital produced by the school is:

- decreasing in performance measure **distortion**, $1 - cos(\theta)$
- decreasing in performance measure **noise**, σ_ϕ^2

Distortion Distortion captures the correlation between the piece rates for different actions and the marginal return to human capital of those actions. In essence, do we pay teachers more for the actions which are more related to developing human capital? The more distorted a contract is, the more employees focus on actions that are less helpful toward firm outcomes.

Noise Noise captures how much of the performance incentive is unrelated to employee's actions. This could operationalize as other factors outside the employee's control affecting the performance measure (school resources, shocks, etc.) or mis-measurement of employee actions, if the contract attempts to measure teacher actions. It is important to flag that traditionally the way noise enters the optimal contract design is through reducing risk-averse employee's utility. This requires firms to raise the fixed part of an employee's salary to meet employee's participation constraint. Here we are not focused on that consequence of noise as we are not focused on employee entry or exit in this paper.[5]

[4] A large class of incentives, including all tournaments, have a fixed variance, so the predictions of the model, apply in those cases as well.

[5] A companion paper [Brown and Andrabi, 2020] studies employee sorting in response to these contracts

The effect of noise we focus on here is equivalent to a decrease in the incentive scheme's average piece rate. Since $\sigma_\phi^2 = var(P) - \|g\|^2 var(\mathbf{a})$, and $var(P)$ and $\mathbf{a}$ are constant given the tournament nature of each incentive scheme, increasing σ_ϕ^2 directly decreases $\|g\|$. Therefore, increasing noise then reduces the extent of the effort response, $\|\mathbf{a}^*\|$. This effect of noise exists in any incentive scheme with a fixed variance, which includes all tournament or threshold-type incentives.

Understanding the Experiment within the Theoretical Framework

The theoretical framework allows us to understand incentive scheme's key features that should affect how teachers respond and, as a result, the impact on human capital. Ex-ante, it is not clear whether subjective or objective incentives would be more or less distorted in the teaching context. On the one hand, subjective incentives may solve the multi-tasking problem by prioritizing more than just measurable student learning. One of the key critiques of objective incentives is that teachers may focus on actions which enhance test scores (such as test prep skills, memorization, etc.), but have small or zero effects on human capital [Muralidharan and Sundararaman, 2011]. Subjective performance incentives would ideally penalize these types of behaviors by teachers, in favor of more well-rounded teaching. On the other hand, it could be that managers prioritize the wrong actions – because they do not know what the human capital production function is, because they value only certain aspects of human capital and not others, or most nefariously, they weight actions which make their job easier.

It is also uncertain whether subjective of objective would be less noisy. Test scores are notoriously noisy measures of teacher effort [Chetty et al., 2014a]. One of the most common complaints teachers have against test score-based incentives is that they are mostly unrelated to teacher actions [Podgursky and Springer, 2007]. Subjective performance pay could be less noisy than objective performance pay because managers could focus on rewarding actions rather than outcomes. However, this requires managers to observe effort accurately. Subjectivity could even introduce additional noise though, if managers introduce bias or favoritism into their evaluations.

Our experiment connects to the model in two ways. First, in sections 2.5, we explicitly test the two predictions of the model using exogenous variation within one of the treatment arms that varies the level of noise and distortion. We then see the effect of these mechanisms on firm outcomes. Second, in section 2.5 and 2.5 we show that the difference in reduced form effects of each contract can be explained through differences in noise and distortion across the two contracts.

2.3 Experimental Design

Performance Incentive Treatments

We partnered with a large private school system in Pakistan to implement the research design. Schools are randomized to receive one of three contracts which determine the size of teachers' raises at the end of the calendar year.[6] The three contracts were:

- **Control: Flat Raise** - Teachers receive a flat raise of 5% of their base salary
- **Treatment: Performance Raise** - Teachers receive a raise from 0-10% based on their within-school performance ranking

Performance Group	Within-School Percentile	Raise amount
Significantly above-average	91-100th	10%
Above-average	61-90th	7%
Average	16-60th	5%
Below average	3-15th	2%
Significantly below average	0-2nd	0%

There are two treatment arms, which vary what performance measure is used to evaluate teachers. Teachers in a given treatment arm are ranked within their school on one of the following performance measures:

  - ***Subjective Treatment Arm:*** Teachers are evaluated by their manager at the end of the calendar year. Managers had complete discretion over how they evaluated teachers and what aspects of performance they would prioritize. To ensure teachers knew what was expected of them, managers delineated between 4-10 evaluation criteria, which would be used to evaluate the teachers. These included items such as improving their behavioral management of students, assisting with administrative tasks, helping plan an afterschool event, and improving students' spoken English proficiency.[7]
  - ***Objective Treatment Arm:*** Teachers are evaluated based on their average percentile value-added [Barlevy and Neal, 2012] for the spring and fall term. Percentile value-added is constructed by calculating students' baseline percentile within the entire school system and then ranking their endline score relative to all other students who were in the same baseline percentile.[8] We then average across all students the teacher taught during the two terms.

[6] Pairwise randomization by baseline test performance was used, which generally performs better than stratification for smaller samples [Bruhn and McKenzie, 2009].

[7] An example set of criteria are provided in Appendix Table A.15.

[8] Percentile value-added has several advantageous theoretical properties [Barlevy and Neal, 2012] and is also more straightforward to explain to teachers than more complicated calculations of value-added.

The contract applied to all core teachers (those teaching Math, Science, English, and Urdu) in grades 4-13. Elective teachers and those teaching younger grades received the status quo contract. All three contracts have equivalent budgetary implications for the school. We over-sampled the number of subjective treatment arm schools due to partner requests, so the ratio of schools is 4:1:1 for subjective treatment, objective treatment, and control, respectively.

Both the subjective and objective treatment arms have several features in common, allowing us to isolate the effect of differing the performance metric and nothing else about the incentive structure. Both treatments are within-school tournaments, so this holds the level of competition fixed between the two treatments. In addition, the variance in the distribution of the incentive pay is equivalent across the two treatments. As we showed in section 2.2, holding the variance constant allows us to interpret differences in noise levels between the two systems as equivalent to differences in incentive steepness. The performance evaluation timeline also played out the same for all groups. Before the start of the year, managers set performance goals for their teachers irrespective of treatment. Teachers were evaluated based on their performance in January through December, with testing conducted in June and January to capture student learning in each term of the year.[9]

To ensure teachers and managers had full understanding of how each contract would work, we conducted an intensive information campaign with schools. First, the research team had an in-person meeting with each manager, explaining the contract assigned to their school, and, in the case of the subjective treatment, explaining what would be expected of them and when. Second, the school system's HR department conducted in-person presentations once a term at each school to explain the contract. Third, teachers received frequent email contact from school system staff reminding them about the contract and half-way through the year contract teachers were provided midterm information about their rank based on the first 6 months.[10] Control teachers were also provided information about their performance in one of the two metrics, in order to hold the provision of performance feedback constant across all teachers.

Timeline and Data

Our study was conducted from October 2017 through June 2019. It covered one performance review cycle conducted from January-December 2018 in which the contracts

[9]The school systems' central office designed and administered the June test to all students in a given grade. However, tests are graded locally by the school, often by the students' teacher. Due to concerns of grade manipulation, grading was audited by the research team. 10% of all teacher's exams were regraded. If the teachers' grade and the auditor's grade were off by more than 5%, another 10% of their tests were audited. If the average was still off by more than 5%, all of the teacher's exams were regraded. Overall, grade manipulation was small and was generally driven by cases where teachers bumped up students' grades from just failing to just passing. There was no heterogeneity in grading accuracy by treatment. The January test was conducted exclusively by the research team (described in section 2.3 below). These tests are *not* used as an outcome measure in this paper.

[10]An example midterm information note is provided in Appendix Figure A.16.

were in place. Figure 2.1 presents the main treatment implementation (detailed in section 2.3) and data collection activities (detailed below).

Our data allows us to understand how teachers changed their effort under each incentive scheme, why the incentives affected effort in the way they did, and the resulting effect this had on student outcomes. We draw on data from (i). the school system's administrative records, (ii). baseline and endline surveys conducted with teachers and managers (iii). endline student testing and survey and (iv). detailed classroom observation data.

Administrative Data: The administrative data details position, salary, performance review score, attendance, and demographics for all employees. We also have biometric clock in/out data for all schools. The data was provided by the school system for the period of July 2016 to June 2019. It includes classes and subjects taught for all teachers, and end of term standardized exam scores for all students (linked to teachers). From September through December 2018, we also have data on classroom observations conducted by managers. Managers use a similar rubric to the one used by the research team to conduct classroom observations (detailed below).

Baseline Survey: The baseline survey measured teachers' preferences over different contracts and beliefs about their performance under each contract. 40% of schools were randomly selected to participate in an in-person baseline survey conducted in October 2017. 2,500 teachers and 119 managers were surveyed. These outcomes are primarily used for a companion paper on teacher selection in response to performance pay [Brown and Andrabi, 2020].

Endline Survey: The teacher endline survey measured their understanding of the contract they were assigned, time use, and beliefs about their manager's level of bias in conducting performance evaluations. The manager endline survey measured managers' beliefs about teacher quality and measured management quality using the World Management Survey school questionnaire.[11] The endline survey was conducted online with teachers and managers in spring and summer 2019. 6,080 teachers and 189 managers were surveyed.

Endline Student Testing and Survey: An endline test was conducted with students to measure performance in core subjects and socio-emotional skills after one year of the intervention. The research team conducted the endline test and student survey in January 2019. The test was conducted in Reading (English and Urdu), Math, Science, and Economics. The items were written in partnership with the school system's curriculum and testing department to ensure appropriateness of question items. Grading was conducted by the research team. Items from international standardized tests (PISA, TIMSS, PERL, and

[11]Due to budget constraints, we were unable to have the World Management Survey surveyors conduct the survey. Instead, we asked managers to directly rate themselves on the rubric that surveyors use. This approach could result in inflated management scores. As a result, we use additional objective data to corroborate the management scores.

LEAPS) and a locally used standardized test (LEAPS) were also included to benchmark student performance.[12]

Students also completed a survey to measure four areas of socio-emotional development. The areas are (i). love of learning (items drawn from National Student Survey, Learning and Study Strategies Inventory), (ii). ethical (items from Eisenberg's Child-Report Sympathy Scale, Bryant's Index of Empathy Measurement), (iii.) global citizen (items from Afrobarometer; World Values Survey), and (iv.) inquisitive (items from Learning and Study Strategies Inventory; Epistemic Curiosity Questionnaire). Appendix table A.14 lists the survey items used for each area along with their source.

The choice of these four areas came from the school system's priorities. They are the four areas of socio-emotional development they expect their teachers to focus on. These areas are posted on the walls in schools, and teachers receive professional development on these areas. Some managers also specifically make these areas part of teachers' evaluation criteria. In addition to these four areas, the survey also asked whether students liked their school or wanted to change to a different school.

Classroom Observation Data: To measure teacher behavior in the classroom, we recorded 6,800 hours of classroom footage and reviewed it using the Classroom Assessment Scoring System, CLASS [Pianta et al., 2012], which measures teacher pedagogy across a dozen dimensions.[13] [14] We also recorded whether teachers conducted any sort of test preparation activity and the language fluency of teachers and students.

Performance Evaluation Data: The school system had an existing performance evaluation system in which managers rated their teachers in December on performance criteria set in the previous December. We layered these new contracts on top of that existing system. In December 2017, before the announcement of treatments, managers set a number of performance criteria for each teacher, as they do each year. In a randomly chosen 3/4 of the subjective schools, those goals then become the evaluation criteria used to determine teachers' raises for the following year. In the rest of the schools (objective, control, and

[12]The endline student test data was used both for evaluating the effect of the treatments and used to compute objective treatment teachers' raises.

[13]There are tradeoffs between conducting in-person observations versus recording the classroom and reviewing the footage. Videotaping was chosen based on pilot data which showed that video-taping was less intrusive than human observation (and hence preferred by teachers). Videotaping was also significantly less expensive and allowed for ongoing measurement of inter-rater reliability (IRR).

[14]We did not hire the Teachstone staff to conduct official CLASS observations as it was cost-prohibitive and we required video reviewers to have Urdu fluency. Instead we used the CLASS training manual and videos to conduct an intensive training with a set of local post-graduate enumerators. The training was conducted over three weeks by Christina Brown and a member of the CERP staff. Before enumerators could begin reviewing data, they were required to achieve an IRR of 0.7 with the practice data. 10% of videos were also double reviewed to ensure a high level of ICC throughout the review process. We have a high degree of confidence in the internal reliability of the classroom observation data, but because this was not conducted by the Teachstone staff, we caution against comparing these CLASS scores to CLASS data from other studies.

the remaining subjective) those goals are used to provide feedback to teachers but have no financial consequence. In the remaining 1/4 of subjective schools, managers were required to create a new set of goals now that they knew there would be financial stakes attached to those goals. They were encouraged to set the goals to be focused on employee effort, rather than employee characteristics, like training or credentials. Since the performance evaluation system exists for all employees, we can use data on what goals were set and the scores on those goals to understand manager priorities and ratings with and without financial stakes tied to the performance rating.

Sample and Characteristics of the Employee-Manager Relationship

Teachers The study was conducted with a large private school system in Pakistan. The student body is from an upper middle-class and upper-class background. School fees are $2,300-$4,300 USD (PPP) per year. Teachers are generally younger and less experienced than their counterparts in the US, though they have similar levels of education. Table 2.1 presents summary statistics of our sample compared to a representative sample of teachers in US [National Center for Education Statistics, 2011]. Our sample is mostly female (80%), young (35 years on average), and inexperienced (5 years on average, but a quarter of teachers are in their first year teaching). All teachers have a BA and 68% have some post-BA credential or degree. Salaries are on average $17,000 USD (PPP).

Managers In order to understand the effects of subjective performance pay, we need to understand who the managers are and what role they play in overseeing teachers. Managers here are either a principal in small schools or a vice principal in larger schools. They are tasked with overseeing the overall operations of the school and managing employees, including teachers and other support staff. Table 2.2 presents information about managerial duties compared to a US sample of principals. Like in the US, our managers are generally older (45 years old), less likely to be female (61%), and more experienced (9.6 years) than teachers. Most were previously teachers and transitioned into an administrative role. Managers spend about a 1/3 of their working hours overseeing their staff – observing classes, providing feedback, meeting with teachers and reviewing lesson plans. The rest of their time is spent on other tasks related to the schools functioning. The distribution of time use is fairly similar to the principals in the US.

However, teachers in our sample spend much more time directly observing teachers. They do about twice the number of classroom observations each year (4.7 versus 2.5 in the US). They also rate themselves higher in most areas of the management survey questions (4.3 versus 2.8 out of 5), including formal evaluation, monitoring and feedback systems for teachers. This is an important difference as these management practices could positively effect the success of the subjective treatment arm, and may help us understand the extent of external validity of these results.

Intervention Fidelity

In this section, we provide evidence to help assuage any concerns about the implementation of the experiment. First, we show balance in baseline covariates. Then, we present information on the attrition rates. Finally, we show teachers and managers have a strong understanding of the incentive schemes. Combined, this evidence suggests the design "worked".

Schools in the two treatment arms and control appear to be balanced along baseline covariates. Appendix Table A.5 compares schools along numerous student and teacher baseline characteristics. Of 27 tests, one is statistically significant at the 10% level and one is statistically significant at the 5% level, no more than we would expect by random chance. Results presented include specifications which control for these few unbalanced variables.

Administrative data is available for all teachers and students who stay employed or enrolled during the year of the intervention. During this time 23% of teachers leave the school system, which is very similar to the historical rate of turnover. 88% of teachers completed the endline survey. While teachers were frequently reminded and encouraged to complete the survey, some chose not to. We do not see differences in these rates by treatment.

Finally, for the endline test, parents were allowed to opt out of having their children tested. Student attrition on the endline test was 13%, with 3 pp of that coming from students absent from school on the day of the test and the remaining 10 pp coming from parents choosing to have students opt out of the exam. On both the endline testing and endline survey, we do not find differences in attrition rate by treatment. We also do not find that lower performing students were more likely to opt out.

Teachers have a decent understanding of their treatment assignment. Six months after the end of the intervention, we ask teachers to explain the key features of their treatment assignment. 60% of teachers could identify the key features of their raise treatment. Finally, most teachers stated that they came to fully understand what was expected of them in their given treatment within four months of the beginning of the information campaign.

2.4 Results

We now present the main reduced form results of the paper. First, we test the effects of each incentive on student test performance and socio-emotional development. Then, we show the effects of the incentives on teacher effort, which helps us to understand the student effects.

Effect of Incentives on Student Outcomes

Specification

Our main specification is:

$$Y_{i1} = \alpha + \beta_1 SubjectiveTreatment_s + \beta_2 ObjectiveTreatment_s + \delta Y_{i0} + \chi_j + \epsilon_i \quad (2.6)$$

The main dependent variable of interest is student outcome, Y_{i1}, for child, i, at endline, t=1. Student outcomes include test scores in Math, Science, English and Urdu and socio-emotional development. $SubjectiveTreatment_s$ and $ObjectiveTreatment_s$ are a dummy for whether the student's school, s, was assigned to subjective or objective performance raises. The left out group is the control group (flat raise). The coefficients of interest are β_1 and β_2, and their test of equality. For test scores, we control for student's baseline score, Y_{i0}, to improve efficiency as there is high auto-correlation in test scores.[15] We also control for strata fixed effects, subject and grade, χ_j. Standard errors are clustered at the school level (the unit of randomization), and both standard and randomization inference p-values are provided in each table.

Results

Test Scores We find that both subjective and objective performance incentives have similar effects on test scores, of about 0.09 sd. Table 2.3 presents the results of each performance incentive on endline test scores. Column (1) shows results for all tests and question items. Effects are similar between the subjective and objective incentives, with an effect of 0.086 sd and 0.092 sd, respectively. In the row titled "F-test p-value (subj=obj)", we present a test for the equality of $\beta_1 = \beta_2$. We cannot reject equality of effects between the two treatments on test scores. All results appear unchanged whether we consider standard p-values (in parentheses) or randomization inference p-values (in brackets).

Column (2) and (3) provide tests on the effect of the treatment by question item type to understand whether these effects are due to memorization of class content or actual learning. Column (2) only includes questions from the prior grade's content and column (3) only includes questions that were added by the researchers from external standardized test sources including PISA, TIMSS, PERL and LEAPS.[16,17] Both sets of questions provide a useful test because it would not be possible for students to have memorized the answers to the questions. Remedial content (from previous grade levels) and external content are never tested on the school system's standardized exam, and so teachers would not have prepared

[15] For grade 4 students we do not have a baseline because standardized testing in the school system begins in 4th grade. For these students, and any other that are missing a baseline, we denote a score of zero and add a dummy for having the baseline test missing.

[16] Not all subject and grade exams had remedial questions or external, so this is reflected in the decrease in sample size.

[17] Question items derived from these international sources were relevant to the curriculum of this school system and were not always matched to corresponding grade from the international exam, if that content was not part of the given year's curriculum.

specifically for this material. Given that we find similar if not larger effects on these types of questions, it appears that treatment effects are coming from actual learning as opposed to memorizing curriculum. Again, we do not see a significant difference between the subjective or objective treatment.

Column (4) and (5) present the results by subject, splitting by math and science exams versus the two reading exams (English and Urdu). Magnitudes are similar, around 0.09 sd, for both subjects, though we are less powered to detect overall effects with the smaller sample when we split by subject. Again, we cannot reject equality between the two treatments and the magnitude of the effects is highly similar.

Socio-Emotional Development While the effects on test scores were similar between both treatments, the effects on socio-emotional development paint a very different picture. Table 2.4 presents the results on socio-emotional development overall and broken down socio-emotional area. Objective incentives result in a small negative effect on socio-emotional development, whereas there is a small positive effect of subjective incentives. When we split these results into their sub-areas, we see that the overall negative effect of objective incentives is coming from a negative effect on "love of learning" and whether students like their school or would like to change schools. We can reject equality of the two treatments on these sub-areas at the 10% and 1% levels, respectively. This suggests that while objective incentives led to an increase in test scores, it was at the cost of enjoying school. Whereas, subjective incentives were able to accomplish the same learning gains without these negative consequences. On three other areas, ethical behavior, being a global citizen and inquisitiveness, we cannot reject the equality of the two treatments.

Effect of Incentives on Employee Effort

Specification

To understand why we see similar results on test scores but different effects on student's socio-emotional development, we need to understand teacher's behavioral response. To do this, we look at the effect of each treatment on classroom observation ratings and time use. We have a similar main specification, this time at the teacher level:

$$Y_i = \alpha + \beta_1 SubjectiveTreatment_s + \beta_2 ObjectiveTreatment_s + \chi_j + \epsilon_i \tag{2.7}$$

The main dependent variable of interest is outcome, Y_i, for teacher, i. Teacher outcomes include classroom observation scores and time use. We again control for grade and strata fixed effects, χ_j, and standard errors are clustered at the school level (the unit of randomization).[18]

[18]We do not control for subject here, unlike in our student specification, because most teachers teach several subjects. In addition, for classroom observations, the observation period often overlapped with several subjects.

Results

Classroom Observations The effect of each incentive on classroom behavior sheds light on the student effects we see. Overall, we find teachers under objective incentives using teaching strategies which provide the largest marginal return on test scores but may hamper other areas of human capital development for students. Teachers in the subjective treatment however, do not exhibit any of those distortionary teaching strategies.

Table 2.5 presents the effects of each incentive on teachers' overall classroom observation score, using the CLASS rubric. On average, objective teachers exhibit worse teaching pedagogy. They score 0.07pts lower on the 7pt CLASS rubric scale. Subjective teachers have no noticeable change in pedagogy quality, and we can reject the equality of the two treatments at the 10% level.

We then break down the 12 CLASS dimensions of pedagogy into three main areas, "class climate", "differentiation", and how "student-centered" the lesson is. "Class climate" captures whether the atmosphere of the classroom is positive, supportive and joyful or negative, punitive and dull. "Differentiation" captures whether the lesson is structured in a way to meet students who are different proficiency levels and/or have different learning styles. Finally, "student-centered" measures how much of the lesson is teacher-directed versus student-involved. Teachers under the objective incentive contract have a more negative class climate and less student-centered lessons. Both see a decrease of around 0.1 pts. We can reject equality of treatments at the 10% level. There is also an increase in level of differentiation in the subjective and objective treatment schools.

We also measure the amount of class time devoted to test preparation activity. This includes practice tests, testing strategies (such as how to approach a multiple-choice test), or lecturing about the importance of doing well on tests. We find a large increase in the time spent on these activities in objective treatment schools. Relative to a control group mean of 0.14 min out of the 20-minute observation spent on test preparation activities, objective classes see a 5-fold increase, with a total of 0.76 minutes spent on these activities. We can reject equality of treatments at the 5% level along this dimension.

Together with the student outcomes, these classroom observations paint a picture of objective schools as ones that were able to achieve test score gains by taking the path of least resistance for teachers – doing more test preparation and maintaining a stricter, less student-centered classroom. This then results in other negative outcomes on students human capital development, such as love of learning. Subjective classrooms on the other hand are able to accomplish the same academic gains without any negative effects on teacher practices or student socio-emotional development. This suggest that managers are able to prevent these distortionary behaviors, solving, at least to some extent, the multi-tasking problem.

One concern with classroom observation data is that teachers may worry the videos of their classrooms will be provided to their manager, and for subjective teachers that has more a consequence than for the other treatment arms. We do several things to help alleviate these concerns. First, in the consent form and during the camera set up, we communicate to teachers that the videos are confidential and will only be reviewed by the research team. We also let them know that only aggregated data at the school level will be provided to the

school system head office. Second, visits were a surprise within a two-month window, so teachers could not adjust their lessons beforehand. Third, we recorded several hours back to back for each teacher. We find teachers are most aware (and responsive) of the camera in the first hour of taping. We can remove that data and repeat the same analysis and find very similar results.

Attendance and Time at Work We find that the subjective treatment results in a significant increase in the number of days a teacher is present at work. Table 2.6 presents the results of the biometric clock in/out data. Relative to a control group mean of 145 days, subjective teachers are present an additional 6 days. We do not find an effect on hours spent at work for either treatment relative to the control. We cannot reject equality of treatments in either outcome. Columns (2) and (4) restrict to a sample of teachers who were present in the school system both terms and did not take any long leaves (health, maternity, etc.) to ensure the days present result is not driven by these effects. Results are robust to this sample restriction.

How do Managers Implement the Subjective Incentive?

In the objective treatment schools there is less scope for heterogeneity. The implementation of the contract and employee's response is likely to be similar across schools and comparable to other experiments which used test score-based performance pay. However, the subjective treatment arm could vary substantially across schools and firms depending on the type of oversight managers have of employees, the oversight firms have on managers and how managers themselves are incentivized.

In this section, we unpack what types of teacher actions managers value, the extent to which managers are biased or show favoritism, and heterogeneity in treatment effects my manager quality. To understand how managers use the subjective treatment arm, we draw on data from the endline teacher and manager survey and managers evaluation scores of their teachers.

What do managers value in rating teachers? We use three approaches to help understand what types of teacher actions' managers reward. In an ideal setting, we would randomize teacher actions to see how this affects managers' performance ratings of teachers. We are unable to do that exact exercise here. However, using a combination of detailed data and survey vignettes, we can accomplish something similar. Combined, these three sources of evidence suggest that managers highly value teacher actions which are related to human capital development and are not just focused on administrative tasks or actions unrelated to student development.

Our first piece of evidence on what managers value in teachers, comes from endline survey data from both teachers and managers. We asked both teachers and managers to respond to a hypothetical situation, in which a teacher asks them for advice about how to achieve a higher raise in the following year. They are then asked to rate how much time the teacher

should spend on different types of actions. Table A.9 presents the data from the survey question. Column 2 shows teachers' responses about which actions would be most highly valued under the subjective contract. Column 3 presents responses to the same question posed to managers. Both subjective teacher and managers agree that improved pedagogy, like making lessons student centered and tailoring lessons to students at different initial levels, would increase their subjective rating. However, managers put additional weight on spending time collaborating with other teachers. Neither subjective teachers nor principals believe more superficial administrative tasks like volunteering at afterschool events or meeting with parents are important drivers of the subjective performance rating.

Our second piece of evidence also comes from the endline survey. We provide a vignette describing a hypothetical teacher to managers, and we ask them to provide a performance rating of the hypothetical teacher. The vignette randomizes the hypothetical teacher's name, and rank in terms of value added, classroom behavioral management and attendance.[19] Table A.10 presents managers' responses to this survey question. We find that managers highly value all three performance characteristics, but place double the weight on teacher value-added as they do on behavioral management and attendance. On average, moving from the 50th percentile value added to the 90th percentile value added would increase a teacher's subjective rating by 0.7sd. Columns 1 through 3 of the table test each attribute separately. Columns 5 and 6 add all attributes together, and we see no difference in relative preference for these teacher characteristics. These results are also robust to adding manager fixed effects.

Finally, we can look at what teacher behaviors are correlated with teachers' actual performance rating in the subjective treatment arm. Table A.11 shows the relationship between teachers' performance rating and teacher behaviors, as measured from classroom observation data, teacher value-added and biometric clock in/out data. We find that managers value higher value added and teacher attendance.[20] This relationship remains when we control for subject and grade (column 2) and classroom observation scores (column 3). We find mixed evidence on the relationship between pedagogy and subjective rating. Some aspects of good pedagogy are valued (teachers who have a negative class climate have a lower rating) but others are not (teachers who spend more time on analysis/inquiry skills and have more student vs teacher talk time are negatively rated). One important limitation with this approach is that there are certainly omitted variables which we are unable to capture. However, having detailed classroom observation and time use data help us paint a relatively detailed picture of each teacher's behavior. Combined these three pieces of evidence suggest that managers have preferences which are relatively aligned with the preferences of the school system.

[19]The vignettes stated, "[Female name/Male name] is in the [bottom/middle/top] 10% of teachers in terms of students' test score growth, in the [bottom/middle/top] 10% of teachers in terms of behavioral management, and is in the [bottom/middle/top]10% in terms of attendance and timeliness at work." Managers rated three such vignettes with characteristics randomized across vignettes.

[20]There is a negative relationship between subjective rating and hours spent at school. This relationship may be driven by the fact that certain grades and teaching positions have different requirements about the length of the workday, so this could be picking up that variation rather than teacher effort.

Favoritism and bias A primary concern about subjective performance pay is whether managers are biased against certain employees or show favoritism toward preferred individuals. To assess whether this is a significant concern in this setting, we ask teachers at endline whether they felt their manager discriminated against certain groups or played favorites toward certain colleagues.[21] Table A.12 presents the results from these survey questions. On average, teachers in the subjective treatment arm are no more likely than teachers in the objective treatment arm to say that the contract unfairly favors certain teachers or that certain groups are discriminated against under this contract. Teachers also state that bias, gaming and favoritism is not a significant concern in either contract.

Though teachers do not say that overt bias is a significant concern, we may be worried that there are more subtle types of bias at play. The primary type of bias we were concerned about in this setting is gender bias. In Pakistan, gender bias in employment is rampant [World Bank Group, 2018], and managers are more likely to be male then the employees they oversee. As part of the vignette survey questions, we include a way to test for subtle gender bias. In the vignettes we randomize the hypothetical teachers' name to be a traditionally male or female Pakistani name. Table A.10, column 3, presents the results of this test. We do not find that managers rate vignettes with female names lower.

Both of these pieces of evidence suggest that favoritism and bias is not a substantial concern within the subjective treatment arm. Neither result is able to perfectly measure whether any favoritism or bias occurred, but combined they provide suggestive evidence favoritism and bias are not a first-order concern under this contract.

Heterogeneity in treatment effects by manager characteristics On average the subjective treatment arm appears to have been successful at improving student outcomes and teacher effort, but there may be heterogeneity in how successfully managers implement the contract. We test for heterogeneity in treatment effects along several dimensions. First, table A.13 presents heterogeneity in the subjective treatment arm by three manager characteristics: gender, age, and experience. We do not find significant differences in the effectiveness of the subjective treatment by these manager characteristics.

Second, table A.13 presents heterogeneity in treatment effects by several dimensions of manager "quality". We find that subjective performance pay is significantly less effective in schools where teachers believe their managers do not have an accurate perception of teacher effort. We measure this by asking teachers to rate how accurate their manager is in rating a fellow teacher.[22] We find there is no effect of subjective performance pay on student test

[21] One concern with this approach is that teachers may be hesitant to provide honest assessment in a survey. To help minimize this concern teachers' responses are anonymized and we communicate this to teachers at the time of consenting to the survey. We also ask the question several ways, including asking teachers to report such behavior about other schools or about the school system in general. This type of questions phrasing allows teachers to report problematic manager behavior while providing plausible deniability for their own manager.

[22] To measure whether a manager has an accurate perception of what their teachers do, we ask teachers to answer the following question about three fellow teachers in their school, "The appraisal score their manager would give them is... [Too high/low by more than one raise category], [Too high/low by about one raise

scores for managers who are in the top quintile of this inaccuracy measure. We do not find heterogeneity and treatment effects by world management survey overall manager score (shown in Table A.13, column (5)) or personnel management sub-score (column 6). However, as discussed in section 2.3, because this data was collected from manager self-report, we should be cautious about the interpretation, as managers may over rate themselves on these survey questions. This suggests that while subjective performance pay is on average very successful at producing learning games, these contracts may be ineffective in settings where employees do not trust their managers to implement them accurately.

2.5 Mechanisms

How can we square the results that we see very different effort responses, similar test score effects and different socio-emotional effects across subjective and objective incentives? We argue that differences in the levels of noise and distortion across the two treatments help explain these outcomes. We structure our argument as follows.

First, in section 2.5, we present the similarities between the two treatments to help eliminate possible channels that could drive the difference in treatment effects. Second, in section 2.5, we highlight the differences between the systems. We show teachers believe subjective incentives to be less noisy and less distorted. Third, we provide evidence that noise and distortion does, in fact, affect outcomes. Section 2.5 shows that noise and distortion are related to student outcomes as predicted in the theoretical framework – more noise reduces the effect of incentives and more distortion diverts employee effort toward those actions. We conduct these tests by exploiting heterogeneity in levels of noise and distortion *within* a given treatment, to isolate the effect of noise or distortion on outcomes. Finally, in section 2.5, we bring together the estimates from section 2.5 and 2.5 to understand how much of the difference in the reduced form student effects can be explained by differences in noise and distortion.

Similarities between Treatments

In order to isolate the effect of the performance measure (percentile value-added versus manager rating), we hold a number of features constant between the two treatments. Both treatments are within-school tournaments. Both treatments provide a raise from 0-10% with the same set of rank thresholds corresponding to raise amounts within that range. Both treatments were introduced at the same time in schools and had a similar performance review timing – manager completed midterm feedback in June 2018 and final ratings in December 2018 and the objective score was based on the average of tests in June 2018 and January 2019.

At endline, we survey teachers about their experience with their incentive scheme. We find no difference in teachers reported experience along a number of dimensions. There is no

category], [Too high/low by less than one raise category], or [Accurate]". We then construct an average of these ratings per manager;

difference in their responses to the following survey questions: i). when teachers said they understood what was expected of them, ii). awareness of contract main features, iii). how frequently they thought about their contract, and iv). whether the system unfairly favors certain types of teachers (age, gender, etc). Table A.12 provides results for each of these survey questions, showing no statistical difference between teachers' responses by treatment.

Differences Across Treatments: Noise and Distortion

In this section through section 2.5, we will focus on two of the remaining differences between the treatments: noise and distortion. As highlighted in the theoretical framework, noise captures the extent to which a teacher's actions affect their incentive payment. Distortion captures the extent to which actions which have the largest marginal return to human capital also are actions which have a higher effective piece rate under the given performance measure. First, we will show that the levels of noise and distortion are different across the treatments.

Noise We measure noise using teacher's perceptions of the noisiness of their incentive treatment.[23] To measure perceived noise, we ask teachers to agree or disagree (on a 5pt scale), whether under their contract, "their raise is out of their control", "those who work harder, earn more" and whether "I feel motivated to work harder". Figure 2.2 presents the average response to each question with 1 being strongly disagree and 5 being strongly agree. We see that teachers in the subjective treatment, feel their raise is more in their control, hard work is rewarded, and they feel more motivated. The average difference is 0.14sd across the three areas, and we can reject equality of treatments for all three questions at the 5% level.

Distortion We measure distortion using endline survey data from teachers. We ask teachers to imagine a teacher who really wants to receive a higher raise at the end of the year and commits to work ten additional hours a week to increase their raise. Then we ask teachers how much of those ten hours should the teacher allocate different activities, such as collaborating with other teachers, incorporating higher order thinking skills into lessons, preparing practice tests, helping with extracurricular activities, etc. We then group these 17 different actions into four categories: administrative tasks (grading, helping with extracurriculars, monitoring duty), professional development (collaboration, training, improved English skills and content knowledge), pedagogy (use of student-centered and differentiated lessons) and test preparation (achieving certain grade targets).

We find that teachers in subjective versus objective schools feel that there are some slight differences in which actions should be prioritized in order to increase their raise. Table 2.7 presents the differences in stated valuation of each area. Overall, teachers think those under

[23] We think this is preferred to using "actual" noise, measured by seeing how predictive teacher's measured behavior is to their raise. Perceived noise is what matters for teacher's behaviors this last year, and there is likely measurement error that is correlated with treatment in measuring "actual" noise.

the subjective contract should prioritize more administrative tasks and slightly less on test preparation. We will show in the next section that these actions have different implications for student outcomes.

Effect of Noise and Distortion on Outcomes

Noise We showed that teachers believe there is less noise in the subjective performance measure. However, we do not know if noise actually reduces the effectiveness of the incentive scheme. We showed that theoretically with a fixed variance incentive scheme, a more noisy incentive scheme leads to a lower power incentive, but there is limited empirical evidence on this effect.

To test whether noise affects outcomes, we exploit heterogeneity *within* the subjective treatment in noisiness. Managers vary in their accuracy of assessing teacher effort. Some managers observe lessons for each of their teachers every week. Others sit down and review paper lesson plans, and some are more hands off. To measure whether a manager has an accurate perception of what their teachers do, we ask teachers to answer the following question about three fellow teachers in their school, "The appraisal score their manager would give them is... [Too high/low by more than one raise category], [Too high/low by about one raise category], [Too high/low by less than one raise category], or [Accurate]". We then construct an average of these ratings per manager, capturing average *perceived* inaccuracy. On average, teachers believe their managers over or under rate their fellow teachers by 0.8 of an appraisal step (out of the five-step system shown in section 2.3. However, there is considerable heterogeneity. Those most inaccurate quintile of managers are perceived to rate other teachers incorrectly by greater than two steps.

More inaccurate managers may be different than their fellow managers in many ways (experience, age, school environment). However, manager accuracy should only affect perceived noisiness of the incentive scheme in subjective treatment schools. In control or objective treatment schools, managers still rate their teachers but have no control over the incentive raise in those schools. Therefore, we use $ManagerAccuracy{*}SubjectiveTreatment$ as the instrument for $Noise$, controlling for $ManagerAccuracy$ and $SubjectiveTreatment$.

We find that $ManagerRatingInaccuracy_j$ significantly predicts teacher's rating of the noisiness of their appraisal system in subjective but not objective/control schools, as we would expect. A 1 sd increase in manager inaccuracy increases beliefs about the noisiness of the contract by 0.1-0.4 sd in subjective schools. Table 2.8 presents the results from the first stage for data at the teacher and student level.[24] Columns (2) and (4) add additional controls, including teacher's beliefs about the preference for different actions ("distortion") and teacher beliefs about other non-noise features of the contract (timing, understanding, etc). The coefficient on $ManagerAccuracy * SubjectiveTreatment$ is very robust to the inclusion of these controls, suggesting that this instrument is picking up difference in noise and not other features of the contract environment.

[24]The first stage in table 2.8 columns (1) and (2) is at the level of the teacher used for the hours results in table 2.9, column (1) and (2). The first stage in columns (3) and (4) is at the level of the student and used for the student test and socio-emotional skills outcomes in table 2.9, column (3)-(6).

To test for the effect of noise on teacher and student outcomes, we use the following two-stage least squares specification:

$$Outcome_{ij} = \alpha_0 + \alpha_1 ManagerRatingInaccuracy_j + \alpha_2 SubjectiveTreat_i + \alpha_3 \widehat{Noise} + \chi_{ij} + \epsilon_{ij} \tag{2.8}$$

where α_3 is the coefficient of interest, $Noise$ is instrumented using $Manager\ Rating\ Inaccuracy_j * SubjectiveTreat_i$. χ_{ij} are controls, such as school and grade and baseline controls when available for a given outcome.

We find that noise significantly reduces the effectiveness of performance incentives (table 2.9). A 1 sd increase in noisiness of the incentive scheme reduces teachers' hours worked by 13.2 hours per week and reduces test scores by 0.175 sd. We do not find an effect of noise on socio-emotional scores. Because our effective first stage has an f-stat of less than 10, we present the AR test p-values which are our preferred test, given that they are robust to weak instruments in the just-identified case. Columns (2), (4), and (6) add in the same additional controls as in table 2.8 for non-noise features of the contract environment. The effect of noise on hours worked and test scores is robust to the addition of those controls.

Distortion Distortion is a measure of how correlated the marginal returns to human capital for different actions are with the effective piece rates for those actions. In order to measure distortion, we therefore need an estimate of marginal returns to different actions. To do this, we again exploit heterogeneity across managers' preferences for different actions, combined with the subjective treatment. The idea behind this strategy is that managers have different preferences for actions – some state they want teachers to focus more on improving their lesson plans, others want teachers to help out more with administrative tasks, etc. We can interact those preferences with subjective treatment status versus objective and control. We can see the effect of preferences toward certain actions on student outcomes.

$$StudentOutcome_i = \alpha_0 + \alpha_1 SubjectiveTreat_i + \sum_{j=1}^{J} \delta_j Points\ on\ Action\ j_i + \sum_{j=1}^{J} \beta_j Points\ on\ Action\ j_i * SubjectiveTreat_i + \chi_{ij} + \epsilon_i \tag{2.9}$$

Here the coefficient of interest is β_j, which gives the effect of manager preference toward certain types of tasks on student outcomes. Actions are grouped into four categories: admin (grading, helping with extracurriculars, monitoring duty), professional development (collaboration, training, improved English), pedagogy (use of student-centered and differentiated lessons), and test prep (achieving certain grade targets). We also add additional controls to capture other features of the contract environment, such as noisiness, understanding of the contract, etc.

We find that several of the action categories are related to student outcomes. Table 2.10 presents the β_j's for each action category. Professional development and test prep actions are positively related to student test scores. However, test prep is negatively related to student socio-emotional scores. These results are robust to the inclusion of additional controls about the contract environment (table 2.10, column (2) and (4)).

Contribution of Noise and Distortion to Reduced Form Effects

Finally, we can pull the results together to understand the extent to which noise and distortion can explain the reduced form results we saw in section 2.4. To do this we decompose the total reduced form effect into the component from noise, distortion and an unexplained component, ϵ:

$$\begin{aligned} dTestScore &= \frac{\partial TestScore}{\partial Noise} * dNoise + \frac{\partial TestScore}{\partial Distortion} * dDistortion + \epsilon \\ -0.006sd &= \quad -0.17 * -0.14sd + \quad -0.03sd \quad + \epsilon \\ \epsilon &= 0.0002sd \end{aligned} \tag{2.10}$$

The overall effect of subjective relative to objective on test scores was close to zero (-0.006sd, from table 2.3). The effect of noise on test scores is -0.17 (table 2.9) and there is 0.14sd less noise in the subjective arm than the objective arm (figure 2.2). For the distortion component, we repeat the same approach for each of the four action categories (admin, professional development, pedagogy and test prep). We take the difference between subjective and objective for each area (table 2.7), multiply each category with the return to preference for that action on test scores (table 2.10) and sum. In total, $\frac{\partial TestScore}{\partial Distortion} * dDistortion$, then is -0.03. Subjective schools put slightly less focus on test scores. Combined, the positive effect of subjective having less noise and the negative effect of them placing less focus on test scores almost cancel each other out. Overall, the remaining unexplained portion, ϵ, is just 0.0002sd, suggesting noise and distortion are effective at explaining the student results.

We can repeat the same approach for socio-emotional skills.

$$\begin{aligned} dSEScore &= \frac{\partial SEScore}{\partial Noise} * dNoise + \frac{\partial SEScore}{\partial Distortion} * dDistortion + \epsilon \\ 0.0433sd &= \quad -0.06 * -0.14sd + \quad 0.011sd \quad + \epsilon \\ \epsilon &= 0.024sd \end{aligned} \tag{2.11}$$

The overall effect of subjective relative to objective on socio-emotional development was 0.0433 sd (table 2.4). The effect of noise on socio-emotional skills is -0.06 and there is -0.14sd less noise in the subjective arm than the objective arm. The subjective teachers focus more on tasks which are related to socio-emotional skills. Overall $\frac{\partial TestScore}{\partial Distortion} * dDistortion$ is 0.011 sd. The remaining unexplained portion is 0.024 sd, or about half of the difference between the subjective and objective treatment. This is perhaps unsurprising given the results

throughout this section. Noise and distortion were much less related to socio-emotional skills than test scores. This could be because there is in fact a weak relationship between them. Alternatively, we may not be as successful at measuring socio-emotional skills and certainly have a harder time capturing what aspects of teacher's behavior is related to developing these skills. Better measurement along these areas is an important area for future work.

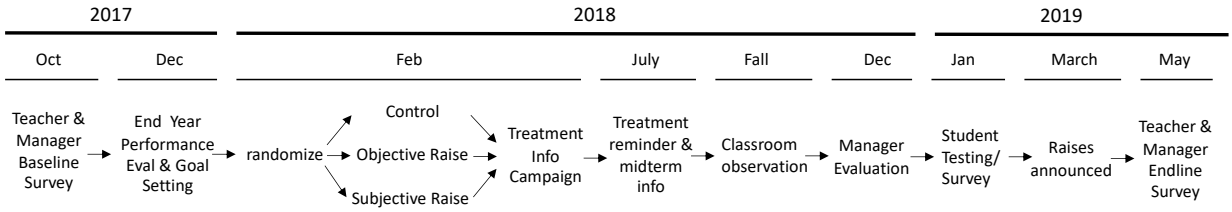
2.6 Conclusion

In this paper, we provide evidence on the effect of subjective versus objective incentives for teachers. We find that both subjective and objective incentives increase test scores, but objective incentives result in negative effects on socio-emotional development. These student outcomes make sense given the teacher behaviors we see under each incentive. In subjective treatment schools, teachers make small improvements in pedagogy and are involved in more professional development. In objective treatment schools, teachers distort effort toward test preparation. They spend much more time on practice tests and test strategies and use more punitive discipline. While there is heterogeneity in manager application of the subjective treatment arm, we do not find evidence of widespread favoritism or bias.

We then try to understand the mechanisms underlying the reduced form effects. We show evidence that the two incentive schemes are similar along most dimensions except for two areas: noise and distortion. We show teachers believe that the subjective incentive is less noisy and that it prioritizes both test and non-test student outcomes. Using heterogeneity within treatments we attempt to isolate the effect of noise and distortion itself on student outcomes. Finally, we show that noise and distortion are able to explain a large portion of the reduced form test score effects but a smaller fraction of the reduced form socio-emotional skill effect.

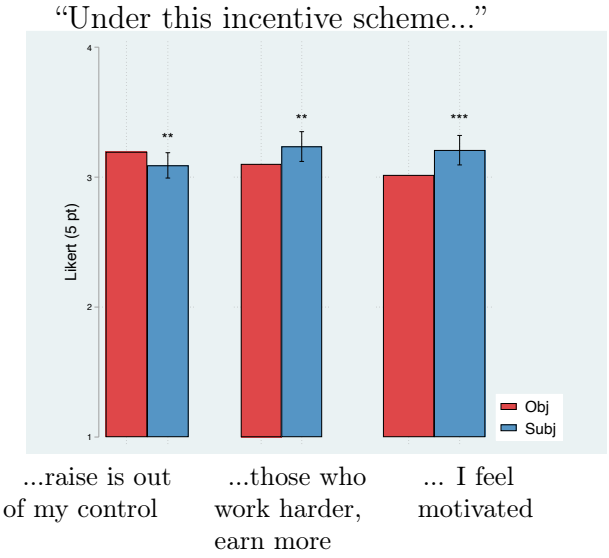
2.7 Figures

Figure 2.1: Experimental Timeline



Notes: This figure presents the experimental timeline. It includes data collection activities and treatment implementation activities.

Figure 2.2: Difference in Noise by Treatment



Notes: This figure presents teacher's responses to questions regarding their incentive contract for the previous year. The question was a on a 5-pt scale from Strongly Disagree (1) to Strongly Agree (5).

2.8 Tables

Table 2.1: Descriptive Statistics about Teachers in Study and Comparison Sample

	Study Sample		US Sample	
	Mean	St. Dev.	Mean	St. Dev.
	(1)	(2)	(3)	(4)
Panel A. Teacher Characteristics				
Age	35.0	8.9	41.8	7.5
Female	0.80	0.40	0.76	0.43
Years of experience	5.1	5.2	13.8	9.6
Has Post BA Education	0.68	0.47	0.54	0.50
Salary, USD(PPP)	17,160	5,700	52,400	18,400
Panel B. Teacher Evaluation				
Number of observations per year	4.7	8.2	2.5	2.9
Use evaluation for compensation	-	-	0.12	0.32
Frequency of evaluation (months)	-	-	13.0	7.0
Performance metric used for evaluation:				
- Principal evaluation	-	-	0.90	0.30
- Test scores	-	-	0.35	0.48
- Peer evaluations	-	-	0.26	0.44
- Student ratings	-	-	0.05	0.22

Notes: This table reports summary statistics on teacher characteristics, monitoring and evaluation for our study sample and a comparison sample of managers in US schools. Data in panel A, columns (1) and (2) comes from administrative data collected from our partner school system. Data in panel B, columns (1) and (2) is from an endline survey conducted with 189 principals and vice principals and 5,698 teachers in our study sample. Data in panel A, B and C, columns (3) and (4) comes from 9,235 principals and 42,020 teachers surveyed in the *School and Staffing Survey* [National Center for Education Statistics, 2011]. Most of panel B is not included for our sample as the experiment determined these features.

Table 2.2: Descriptive Statistics about Mangers in Study and Comparison Sample

	Study Sample		US Sample	
	Mean	St. Dev.	Mean	St. Dev.
	(1)	(2)	(3)	(4)
Panel A. Manager Characteristics				
Age	44.9	9.2	48.8	9.7
Female	0.61	0.49	0.53	0.50
Years of experience	9.6	7.9	13.0	7.5
Salary, USD(PPP)	45,400	34,400	85,400	29,400
Panel B. Manager Time Use				
Total hours worked	47.2	16.3	57.0	13.2
Hours spent on:				
- Administrative tasks	18.5	10.3	18.2	2.3
- Teacher management and teaching	17.5	8.2	15.1	2.0
- Student and parent interactions	6.3	4.4	20.2	2.7
- Other tasks	6.9	12.3	4.0	2.6
Panel C. Management Practice Rating				
Overall Management Score (out of 5)	4.27	0.43	2.76	0.43
People management (out of 5)	4.14	0.53	2.51	0.49
Operations (out of 5)	4.32	0.61	2.89	0.49
Performance monitoring (out of 5)	4.32	0.49	2.81	0.75

Notes: This table reports summary statistics on manager characteristics, time use and management practices for our study sample and a comparison sample of managers in US schools. Data in panel A, columns (1) and (2) comes from administrative data collected from our partner school system. Data in panel B and C, columns (1) and (2) is from an endline survey conducted with 189 principals and vice principals in our study sample. Data in panel A and B, columns (3) and (4) comes from 9235 principals surveyed in the *School and Staffing Survey* [National Center for Education Statistics, 2011]. Data in panel C, columns (3) and (4) is from the *World Management Survey* data conducted by the Centre for Economic Performance [Bloom et al., 2015]. We restrict to the 270 schools located in the US from that sample.

Table 2.3: Effect of Incentives on Student Test Scores

	Endline Test (z-score)				
	All	Remedial	External	Math/Science	English/Urdu
	(1)	(2)	(3)	(4)	(5)
Objective Treatment	0.0918*	0.189***	0.119**	0.104*	0.0917
	(0.0575)	(0.00518)	(0.0335)	(0.0668)	(0.166)
	[0.0730]	[0.0260]	[0.0200]	[0.194]	[0.144]
Subjective Treatment	0.0859**	0.142**	0.0855*	0.0884*	0.0986**
	(0.0220)	(0.0113)	(0.0601)	(0.0646)	(0.0267)
	[0.0130]	[0.0240]	[0.0170]	[0.121]	[0.0260]
F-test pval (subj=obj)	0.89	0.38	0.43	0.77	0.90
Randomiz infer pval (subj=obj)	0.884	0.453	0.388	0.819	0.873
Control Group Mean	-0.04	-0.09	-0.05	-0.04	-0.04
Clusters	234	204	225	223	225
Observations	141566	31944	100318	72714	68852

Notes: This table presents the effects of each performance incentive treatment on student endline test scores. The outcome is student's z-score on a given endline exam. The sample includes students tested in grades 4-13 in five subjects: Math, Science, English, Urdu, Economics. Column (1) includes all test subjects and question items. The observation is at the student-subject exam level. Column (2) restricts to question items which were from the previous grade. Column (3) restricts to question items drawn from external sources, such as PISA and TIMSS. Column (4) restricts to math and science exams. Column (5) restricts to English, Urdu and Economics exams. All regressions include strata fixed effects and control for baseline student average test score, baseline school average test score, grade and subject. Values in parentheses are standard p-values. Values in brackets are randomization inference p-values. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table 2.4: Effect of Incentives on Student Socio-Emotional Outcomes

	Socio-Emotional Indices (z-score)					
	All	Love of learning	Ethical	Global	Inquisitive	Dislike school
	(1)	(2)	(3)	(4)	(5)	(6)
Objective Treatment	-0.0262	-0.0854	-0.0137	0.0278	0.00293	0.0860*
	(0.423)	(0.133)	(0.760)	(0.582)	(0.955)	(0.0719)
	[0.515]	[0.123]	[0.830]	[0.635]	[0.957]	[0.135]
Subjective Treatment	0.0171	0.000933	0.0115	0.0474	-0.0217	-0.0314
	(0.363)	(0.976)	(0.668)	(0.192)	(0.552)	(0.395)
	[0.576]	[0.985]	[0.792]	[0.225]	[0.649]	[0.513]
F-test pval (subj=obj)	0.16	0.09	0.55	0.65	0.59	0.00
Randomiz infer pval (subj=obj)	0.146	0.0420	0.626	0.682	0.614	0.00400
Control Group Mean	-0.00	-0.00	-0.00	0.00	-0.01	0.38
Clusters	126	126	126	125	126	124
Observations	15418	15401	14904	14168	14909	11505

Notes: This table presents the effects of each performance incentive treatment on student socio-emotional outcomes. The outcome is student's z-score on a given socio-emotional dimension. Observations are at the student level and come from an endline survey of students in January 2019. Column (1) provides the average across all five dimensions of socio-emotional outcomes. Columns (2)-(6) provide each individual dimension. All regressions include strata fixed effects and control for student's grade. Values in parentheses are standard p-values. Values in brackets are randomization inference p-values. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table 2.5: Effect of Incentives on Teacher Effort

	Classroom Observation Rubric				Test Prep
	All	Class Climate	Differentiation	Student-Centered	Minutes
	(1)	(2)	(3)	(4)	(5)
Objective Treatment	-0.0713	-0.0791*	0.110*	-0.115**	0.577***
	(0.123)	(0.0788)	(0.0719)	(0.0346)	(0.00455)
	[0.171]	[0.101]	[0.149]	[0.0480]	[0.0120]
Subjective Treatment	-0.00206	-0.00704	0.105*	-0.0276	0.110
	(0.959)	(0.822)	(0.0699)	(0.521)	(0.255)
	[0.946]	[0.838]	[0.0690]	[0.559]	[0.649]
F-test pval (subj=obj)	0.10	0.10	0.93	0.09	0.02
Randomiz infer pval (subj=obj)	0.109	0.0830	0.940	0.0940	0.0140
Control Group Mean	4.67	5.64	2.65	4.93	0.14
Clusters	142	142	142	142	142
Observations	6827	6827	6827	6827	6827

Notes: This table presents the effects of each performance incentive treatment on teacher behavior as rated based on classroom videos. The unit of observation is at the classroom observation level. Teachers may be observed multiple times over the course of the intervention. Column (1) presents the average score on the CLASS rubric [Pianta et al., 2012], on a 7-pt scale. Columns (2)-(4) provide scores on sub-areas of the class rubric. Column (5) provides the number of minutes during the observation that were spent on testing or test-prep activities. All regressions include strata fixed effects and control for grade and video coder fixed effects. Values in parentheses are standard p-values. Values in brackets are randomization inference p-values. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table 2.6: Effect of Teacher Time at Work

	Days present at school		Hours worked per day	
	(1)	(2)	(3)	(4)
Objective Treatment	2.426	1.554	0.262	0.293
	(0.570)	(0.339)	(0.195)	(0.282)
	[0.618]	[0.392]	[0.318]	[0.319]
Subjective Treatment	5.927*	3.340***	0.0348	-0.0432
	(0.0719)	(0.00947)	(0.840)	(0.832)
	[0.0960]	[0.0100]	[0.855]	[0.823]
Sample	All	Restricted	All	Restricted
F-test pval (subj=obj)	0.30	0.15	0.13	0.12
Randomiz infer pval (subj=obj)	0.371	0.202	0.295	0.164
Control Group Mean	144.79	182.72	7.90	7.92
Clusters	295	277	295	277
Observations	6394	4363	6394	4363

Notes: This table presents the effects of each performance incentive treatment on teacher attendance and time at work. The outcome is the number of days present at work and the number of hours at work. Data comes from biometric clock in and out data collected at all schools. The restricted sample removes teachers who took long leaves of absence or only worked at the school system for one of the two terms. All regressions include strata fixed effects and control for baseline school average test score, grade and subject. Values in parentheses are standard p-values. Values in brackets are randomization inference p-values. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table 2.7: Teachers Perceptions about which Actions to Focus on by Treatment

	Admin (1)	Pedagogy (2)	Prof. Develop. (3)	Test Prep (4)
Subjective Treatment	0.0887*	-0.0175	-0.0513	-0.0623
	(0.0502)	(0.0498)	(0.0495)	(0.0497)
Observations	2887	2887	2887	2887

Notes: This table reports teachers' responses to a hypothetical scenario in which they are advising a teacher which actions they should take to increase their raise under a given treatment. Data was collected as part of the endline survey, and observations are at the unit of the teacher. Actions are categorized into four categories: administrative tasks, pedagogy, professional development, and test preparation. Table A.9 provides teacher's weight for the full list of activities by treatment. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table 2.8: Instrumenting Noise with Manager Accuracy - First Stage

	Noise Index (z-score)			
	(1)	(2)	(3)	(4)
Manager rating inaccuracy (z-score)	0.133***	0.123**	-0.316*	-0.219**
	(0.0502)	(0.0502)	(0.165)	(0.106)
Subjective Treatment	-0.326***	-0.116	-0.887***	0.795
	(0.0626)	(0.0867)	(0.180)	(0.528)
Subjective Treatment*Manager rating inaccuracy (z-score)	0.102*	0.103*	0.419**	0.306**
	(0.0537)	(0.0537)	(0.178)	(0.120)
Sample	Teacher	Teacher	Student	Student
Distortion Controls		X		X
Control Group Mean	0.32	0.32	1.23	1.23
Clusters	290	290	245	245
Observations	3356	3356	436740	436740

Notes: This table presents the relationship between manager rating inaccuracy and teacher's rating of how noisy their contract was. The outcome is teacher's rating of how noisy their contract was as measured by an index of their response to the three questions shown in Figure 2.2. Columns (1) and (2) uses data at the teacher level. Columns (3) and (4) uses data at the teacher-student exam level. Student exam data is matched to all teachers who taught the student in the given exam subject for at least one term from January-December 2018. All regressions control for subject, class and manager inaccuracy squared. Columns (3) and (4) also control for school and student test baseline. Columns (2) and (4) add in additional controls to pick up other non-noise differences across contracts. These controls include weight placed on each of the four activity groups listed in Table 2.7, those values interacted with the Subjective treatment, when teachers said they learned about the treatment and how often they received information about the treatment. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table 2.9: Effect of Noise on Outcomes

	Hours worked		Test Score		Socio-Emotional Score	
	(1)	(2)	(3)	(4)	(5)	(6)
Noise index (z-score)	-13.24	-12.93	-0.175**	-0.269**	-0.0591	-0.211
	(10.10)	(9.982)	(0.0875)	(0.121)	(0.162)	(0.628)
Distortion Controls		X		X		X
AR test p value	0.05	0.06	0.05	0.02	0.63	0.37
Montiel-Pflueger effective first stage F stat	3.60	3.65	5.50	6.46	0.47	0.16
Control Group Mean	40.46	40.46	-0.02	-0.02	-0.00	-0.00
Clusters	290	290	245	245	156	156
Observations	3356	3356	436740	436740	15285	15285

Notes: This table presents the relationship between teacher's rating of the noisiness of their contract, instrumented by manager inaccuracy*Subjective Treatment, on teacher and student outcomes. Columns (1) and (2) use data at the teacher level. Columns (3) and (4) use data at the teacher-student exam level. Student exam data is matched to all teachers who taught the student in the given exam subject for at least one term from January-December 2018. Columns (5) and (6) uses data the student level. All regressions control for subject, class, subjective treatment, manager inaccuracy, and manager inaccuracy squared. Columns (3) and (4) also control for school and student test baseline. Columns (2), (4) and (6) add in additional controls to pick up other non-noise differences across contracts. These controls include weight placed on each of the four activity groups listed in Table 2.7, those values interacted with the Subjective treatment, when teachers said they learned about the treatment and how often they received information about the treatment. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table 2.10: Effect of Manager Preferences on Student Outcomes

	Test Scores		Socio-Emotional Score	
	(1)	(2)	(3)	(4)
Admin*Subjective Treat	-0.134	-0.122	0.00176	0.00823
	(0.103)	(0.101)	(0.0346)	(0.0359)
Professional Development*Subjective Treat	0.248**	0.250**	0.0123	0.0139
	(0.102)	(0.102)	(0.0469)	(0.0474)
Pedagogy*Subjective Treat	0.0394	0.0521	-0.0267	-0.0262
	(0.0892)	(0.0867)	(0.0369)	(0.0368)
Test Prep*Subjective Treat	0.189**	0.190**	-0.163*	-0.162*
	(0.0850)	(0.0853)	(0.0854)	(0.0854)
Noise Controls		X		X
Control Group Mean	-0.02	-0.02	-0.02	-0.02
Observations	2891	2891	2653	2653
Clusters	152	152	100	100

Notes: This table presents the relationship between evaluation criteria interacted with treatment on student outcomes. Data is at the teacher level. All regressions control for the four categories of evaluation criteria and subjective treatment. Columns (2) and (4) add in additional controls to pick up other non-distortion differences across contracts. These controls include noise index, belief about whether the contract affects teacher competition, favors certain teachers, when teachers said they learned about the treatment, how often they received information about the treatment and all of these outcomes interacted with subjective treatment. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Chapter 3

Understanding Gender Discrimination by Managers

3.1 Introduction

Pakistan ranks in the lowest decile in female labor force participation, and even in sectors where women are more prevalent, such as teaching, they earn 70 cents for each dollar men earn. Many papers have documented the extent of gender bias in hiring, wage setting, and promotions across many countries. However, we know much less about the mechanisms explaining these gaps. Is this a result of taste-based discrimination, statistical discrimination, or something else?

In this paper, we use a large-scale experiment with 200 managers and 3,600 employees to test for three potential mechanisms that could explain the gender discrimination in wages we see: taste-based discrimination, statistical discrimination, and, what we call, financial discrimination (or that managers give lower wages to female workers they gauge as equally capable because women are not primary earners). We use data on employees end of year performance evaluations conducted by their manager, along with detailed administrative records on employee attendance and teaching performance. We also draw on endline survey data with managers to gauge their beliefs about their employees' productivity.

To test for taste-based discrimination, we use two sources of data. First, we have managers respond to three questions from the World Values Survey meant to gauge beliefs about women in the workplace. Second, we ask managers to give a hypothetical performance evaluation score for a series of vignettes describing example teachers and their performance, randomly varying whether we use traditionally female or male names in the vignette.

To test for statistical discrimination, we randomly vary how much exposure managers have to certain employees. We do this by requiring managers do additional classroom observations for a randomly selected subset of teachers. Doing these additional observations leads to managers being much better at ranking the teacher's quality on a variety of performance dimensions. Finally, to test for financial discrimination, we randomly vary whether employee's performance evaluation is used to determine their raise at the end of

the year or if a different metric is used to determine their raise. This allows us to test if managers differentially rate female workers lower when there is a financial stake of the evaluation.

First, we find little evidence of taste-based discrimination by managers. Our sample of managers generally has more gender progressive beliefs than the modal individual in OECD countries, which is not too surprising given our managers are highly educated and working in a female-dominated profession. We also do not find that vignettes with female names receive lower scores from managers. As a result of this first finding, we then write a model in which managers do not have taste-based discrimination but potentially suffer from other types of discrimination.

Next, we do find evidence of statistical and financial discrimination by managers. We find that when performance evaluations are tied to the employee's end-of-year raise, female employees receive much lower raises than men (14%), but this is not the case when there is no financial stake associated with evaluation. However, when managers are required to spend more time observing employees, this significantly lowers the extent of gender bias. These effects are consistent across both male and female managers and among managers with low levels of stated gender bias as measured using World Values Survey questions.

3.2 Context

Gender and the Pakistani Labor Market

Pakistan ranks 176 out of 187 countries in terms of female labor force participation, with just 22% percent of women in the labor force. While other metrics of equality, such as education and health, have continued to improve, female labor force participation has only increased 1 percentage point in the last ten years. One acceptable area of the labor force for women to work in is education. In Pakistan, 55% of teachers are women. This gender imbalance is stronger in private schools, such as our setting where 81% of teachers are female. In addition, 61% of our managers are women. Despite the prevalence of women in this sector, they still only earn 70% of what male teachers earn.

This study works with a network of private schools in Pakistan, so our employees are teachers, and the managers are administrators, like principals and vice-principals. Because our sample is from a female-dominated sector and is significantly more educated, we actually find relatively progressive attitudes about gender and employment among our sample of managers. During the endline survey, we ask managers to agree to disagree with several statements used in the World Values Survey to gauge beliefs around women in the workplace. On all three statements, our sample (column 1 of table 3.1) rates as more gender progressive than the average respondent in OECD countries (column 3). These patterns hold for both the male and female managers in our sample. South Asia, in general, though, rates significantly more conservative and is the most conservative region in the world.

However, we may be concerned that individuals do not feel comfortable sharing their views truthfully in the survey as they know it is conducted by a foreign research team. In

order to perform a more naturalistic task that is less blatantly about gender, we provide a series of vignettes to managers about hypothetical employees and ask them to give a performance evaluation score of that employee. In the vignette, we randomly vary the teacher's name (traditionally male versus traditionally female) and the description the teacher's percentile rank in terms of value-added, behavioral management and attendance. The task is framed to the managers as trying to gauge what aspects of teaching managers value, with nothing about gender mentioned during the task. Managers repeat the task for several vignettes with different characteristics.

Overall, we find no relationship between whether a female name is listed and the manager's rating. Column (1) of table 3.2 presents the rating managers give the teacher in the vignette. Column (2) adds in controlling for the other performance characteristics included in the vignette, and column (3) controls for those characteristics interacted with gender. In all three specifications, the coefficient on female name is small and insignificant, and we can reject effects of a negative bias against women of greater than 2.5 percentile or a positive bias toward women of greater than 1.6 percentile at the 5% level. In contrast, we find a very strong relationship between the manager's rating and other performance characteristics like value-added and attendance.

We can also examine if there are differences in the extent of bias on this task by other manager characteristics. Table 3.3 shows the same specification as in table 3.2, column (2) but includes an interaction with various manager characteristic's such as the manager's gender, age, extent of bias on the World Values Survey questions. We find that managers who have higher levels of gender bias on these questions are more likely to rate female named vignettes lower. In particular, responding that you agree or strongly agree that "In general, it is better for a family if a woman has the main responsibility for taking care of the home and children rather than a man." is associated with lower scores for female-named teachers. However, it is only a very small fraction of managers which hold such views.

Combined, the survey and vignette evidence suggests, on the face of it, our sample of managers does not hold discriminatory views of women in the workplace, except for a small fraction of managers. However, as we show in the next section's framework, even without traditional taste-based discrimination by managers, we can end up with disparate employment outcomes by gender.

A Framework for Gender Discrimination

In this section, we write a model to describe wages paid to men versus women in a setting where there are no mean or variance differences in employee quality by gender, and there is no taste-based gender discrimination by managers. The difference in wages by gender and the resulting responsiveness to the policy changes we test will arise from two key features of the model: i). managers have imperfect information about worker quality, and they rely partly on previous performance evaluation scores and ii). workers will complain if performance evaluations are significantly lower than their belief about their quality, and, in this model, men are on average more likely to complain. While removing typical statistical discrimination and taste-based discrimination may underestimate the extent of gender discrimination in

many labor markets, this model helps demonstrate how differences across groups can arise even in settings where people have "good" intentions.

This model describes the performance evaluation process. First, managers rate their employees, then an oversight body reviews the evaluations, and, finally, the employees can complain about the results. Managers take into account the costs of: i). punishment from the oversight body in the event their evaluation is deemed incorrect, and ii). inconvenience if an employee complains about the evaluation. These two factors influence their evaluation decision.

Manager's Decision about Evaluation Score The manager's evaluation score of employee, i, is s_i, which is a function of last year's score, s_{i0}, a performance signal they receive this year (such as the worker's output), y_i, and a component, d_i, which is up to the manager's discretion. The signal of performance is a noisy function of the employee's true productivity, p_i, so $y_i = p_i + \eta_i$, where $\eta_i \sim \mathcal{N}(0, \sigma_\eta^2)$. Employees know their own true productivity, p_i, but cannot credibly signal it to their employer. The evaluation score then is:

$$s_i = (1 - \lambda_i)s_{i0} + \lambda_i y_i + d_i \tag{3.1}$$

where λ_i captures how much weight the manager puts on this year's performance signal versus previous performance scores, which depends on how noisy this year's performance signal is.

The manager chooses the d_i that minimizes the costs from the risk of punishment and the inconvenience cost of an employee complaining. The risk of punishment depends on how much discretion the manager used beyond relying on this year's performance signal and last year's score. The inconvenience cost depends on the difference between the true performance and the evaluation score and a constant, c_g, which can be different for different groups of employees. The manager's utility then is:

$$u(d_i) = \min_{d_i} d_i^2 + c_g(p_i - s_i(d_i)) \tag{3.2}$$

$$\frac{\partial u_i}{\partial d_i} = 2d_i^* - c_g = 0$$

$$d_i^* = \frac{c_g}{2} \tag{3.3}$$

This then implies that the evaluation score the manager will choose is:

$$s_i^* = (1 - \lambda_i)s_{i0} + \lambda_i y_i + \frac{c_g}{2} \tag{3.4}$$

Consistent with the literature on bargaining and negotiation, we assume that men are more likely to contest a low evaluation score, so $c_m > c_f$. This then implies that $s_m^* > s_f^*$. This is the case even with the same distribution in worker productivity, p_i, and noisiness in performance signals, σ_η^2, between men and women.

Comparative Statics The key comparative statics we are interested in are how scores (and the gender gap in scores) respond to changes in the accuracy of information managers have (λ) and the likelihood of complaining for employees (c_g).

$$\frac{\partial s_i^*}{\partial c_g} = \frac{1}{2} \qquad \frac{\partial s_i^*}{\partial \lambda} = -s_{i0} + y_i \tag{3.5}$$

As we would expect, the higher the likelihood an employee complains, the higher their evaluation score. Similarly, the less noisy the performance signal y_i is the more managers will rely on that rather than previous evaluation scores.

The difference in expected scores by gender then is:

$$\frac{\partial s_i^*}{\partial female} = (1-\lambda)(E[s_{i0}|g = female] - E[s_{i0}|g = male]) + \frac{c_f - c_m}{2} < 0 \tag{3.6}$$

There will be a difference by gender in previous evaluation scores and also a difference in the likelihood of complaining, which affects the discretionary component of the score. Finally, what happens to the gender gap when we vary c_g or λ:

$$\frac{\partial^2 s_i^*}{\partial (c_f - c_m)\partial female} = \frac{1}{2} > 0 \tag{3.7}$$

$$\frac{\partial^2 s_i^*}{\partial \lambda \partial female} = -(E[s_{i0}|g = female] - E[s_{i0}|g = male]) > 0 \tag{3.8}$$

Connection to Experiment The focus of the experiment is to test these last two comparative statics. What happens when you reduce the difference in likelihood of complaining (by lowering the financial stakes associated with complaining) and what happens when you improve the accuracy of information employers have about employee quality. We will show evidence that the extent of gender bias in manager's evaluation increases with financial stakes ($\frac{\partial^2 s_i^*}{\partial (c_f - c_m)\partial female} > 0$) and decreases with better information about employee quality ($\frac{\partial^2 s_i^*}{\partial \lambda \partial female} > 0$).

3.3 Experimental Design

Intervention

In order to test our hypotheses, we introduce two variations in how managers within the school system evaluate their employees. The study was conducted from October 2017 to June 2019 with a private school chain that operates nearly 300 schools located across Pakistan. Figure 1.1 presents the timeline of interventions and data collection activities.

Performance Evaluation Cycle In all schools, employees receive an annual performance review. At the beginning of the year, managers sit down with employees and talk about goals and areas for growth in the coming year. Together they make a list of performance areas in which the employee will be evaluated. The employee is also evaluated along fifteen additional criteria which are standard across all individuals with the same job title (teachers, administrators, support staff, etc). For teachers, these criteria range from subject knowledge to interaction with parents. The criteria are listed in the employee's work dashboard and accessible at any point in the year.

Throughout the year, managers are expected to observe the employee's work. In the case of classroom teachers, this takes the form of observing classes, reviewing lesson plans and reviewing graded materials. On average managers observe teachers five times per year. However, there is variation with some managers doing more frequent observations. New hires and less experienced teachers also generally receive more frequent observations.

At the end of the year, managers score employees along the criteria. Employees' total score across all criteria ranges from 0 to 100. Managers are required to give a certain number of employees a score from 90-100, 80-89, and so on. This forced distribution prevents managers from giving everyone very high scores. The table below shows the percent of employees that can fall into each point category Scores are reviewed by the regional offices to ensure some outside oversight on the evaluation. Performance evaluation scores are a permanent part of the employee's personnel records and are accessible to the employee and the employee's supervisors. If the employee changes position or school within the system, their records carry over.

Performance Group	Points	Percent of employees
Significantly above-average	90-100	10%
Above-average	80-89	30%
Average	60-79	45%
Below average	50-59	13%
Significantly below average	Below 50	2%

Once scores are finalized, managers sit down with the employee to discuss their score and provide feedback on the performance in the previous year. They also generally discuss areas to work on improvement in the future. Most employees find the performance evaluation process helpful and constructive.

Treatment 1: Financial Stakes of Performance Evaluation To understand if managers change their evaluations of employees when there are financial stakes, I vary whether manager's end of year evaluation of employees is used just for feedback or if the evaluation also determines the employee's raise at the end of the year.

- **Control:** In control schools, managers complete the performance evaluation cycle as described above. Employee's end of year raise is then determined one of two ways:

  - *Flat Raise:* Employees receive a raise of 5% of their base salary
  - *Objective Raise:* Teachers receive a raise from 0-10% based on their within-school percentile value-added [Barlevy and Neal, 2012] averaged across all students they taught during the spring and fall term exams.[1]

- **Treatment: Subjective Raise:** Teachers receive a raise from 0-10% based on their performance evaluation score.

Under both the subjective and objective raise schools, there is the same distribution of raise values. The top 10% of teachers receive a raise of 10%, the next 30% receive a raise of 7%, the next 45% receive 5%, the next 13% receive 2%, and the lowest 2% of performers receive no raise. The difference is whether their performance evaluation score or their percentile value-added based on end of term test scores is the performance metric used to determine the raise.

Randomization was conducted at the school level, so all teachers at the school were under the same type of raise system.[2,3] The contract applied to all core teachers (those teaching Math, Science, English, Urdu, and Social Studies) in grades 4-13. Elective teachers and those teaching younger grades received the status quo contract. All three contracts have equivalent budgetary implications for the school. I over-sampled the number of subjective treatment arm schools due to partner requests, so the ratio of schools is 4:1:1 for subjective treatment, objective treatment, and control, respectively.

Treatment 2: Increased Observation of Employee Effort In addition to the variation in financial stakes of the performance evaluation, I vary how often managers conduct classroom observations of certain teachers. At the beginning of the second semester of the intervention year, all managers receive a training from the school system on how to use a new classroom observation tool to record their notes and feedback during classroom observations. They are then told they must use the observation tool at least once a month with a randomly sampled set of teachers within their school. For the other teachers, they are allowed to continue their regular frequency of observations. Randomization is at the teacher level, stratified by school. This treatment results in treated teachers receiving a 50% increase in observations in the three-month period before the evaluation scores were due.

[1] Percentile value-added is constructed by calculating students' baseline percentile within the entire school system and then ranking their endline score relative to all other students who were in the same baseline percentile. Percentile value-added has several advantageous theoretical properties [Barlevy and Neal, 2012] and is also more straightforward to explain to teachers than more complicated calculations of value-added.

[2] Triplet-wise randomization by baseline test performance was used, which generally performs better than stratification for smaller samples [Bruhn and McKenzie, 2009].

[3] To ensure teachers fully understood their contract, we conducted an intensive information campaign with schools. First, the research team had an in-person meeting with each principal, explaining the contract assigned to their school. Second, the school system's HR department conducted in-person presentations once a term at each school to explain the contract. Third, teachers received frequent email contact from school system staff, reminding them about the contract, and half-way through the year, teachers were provided midterm information about their rank based on the first six months. An example midterm information note is provided in appendix figure A.16.

Data

We draw on data from (i). the school system's administrative records, and (ii). baseline and endline surveys conducted with teachers and principals.

Administrative data The administrative data details employee job description, salary, performance evaluation total score, score on each performance criteria, attendance, and demographics for July 2015 to June 2019. It includes classes and subjects taught for all teachers and end of term standardized exam scores for all students (linked to teachers).

Teacher and manager survey At baseline and endline, we measure teacher's contract preferences, beliefs about their value-added, and risk preferences. We also conduct a time use survey to understand how much time teachers spend on lesson planning, helping with administrative tasks. The survey also included measures of intrinsic motivation [Ashraf et al., 2020], efficacy [Burrell, 1994], and checks on what teachers understood about their assigned contract. The endline survey was conducted online with teachers and managers in spring and summer 2019. Appendix table A.15 lists the survey items used for each area along with their source.

The manager baseline and endline survey measured managers' beliefs about teacher quality, and the endline measured management quality using the World Management Survey school questionnaire.[4] We measure managers' beliefs about gender roles using questions from the World Values Survey and conduct an exercise to assess principals valuation of different teaching characteristics using sample vignettes.

Sample and Intervention Fidelity

Employees The study was conducted with a large, high fee private school system in Pakistan. The student body is from an upper middle-class and upper-class background. School fees are $900 USD. Table 1.1, panel A, presents summary statistics for our sample teachers compared to a representative sample of teachers in Punjab, Pakistan [Bau and Das, 2020]. Our sample is mostly female (81%), young (35 years on average), and the median experience level is 10 years, but a quarter of teachers are in their first year teaching. Nearly all teachers have a BA, and 68% have some post-BA credential or degree. Teachers are generally younger and less experienced than their counterparts in public schools, though they have more education.

Managers Managers here are either a principal in small schools or a vice principal in larger schools. They are tasked with overseeing the overall operations of the school and managing employees, including teachers and other support staff. Table 2.2 presents information about

[4]Due to budget constraints, we were unable to have the World Management Survey research team conduct the survey. Instead, we asked managers to rate themselves on the rubric. This approach could result in inflated management scores. As a result, we use additional objective data to corroborate the management scores.

managerial duties compared to a US sample of principals. Like in the US, our managers are generally older (45 years old), less likely to be female (61%), and more experienced (9.6 years) than teachers. Most were previously teachers and transitioned into an administrative role. Managers spend about a 1/3 of their working hours overseeing their staff – observing classes, providing feedback, meeting with teachers, and reviewing lesson plans. The rest of their time is spent on other tasks related to the schools functioning. The distribution of time use is fairly similar to the principals in the US.

However, managers in our sample spend much more time directly observing teachers. They do about twice the number of classroom observations each year (4.7 versus 2.5 in the US). They also rate themselves higher in most areas of the management survey questions (4.3 versus 2.8 out of 5), including formal evaluation, monitoring, and feedback systems for teachers. This is an important difference as these management practices could positively effect the success of the subjective treatment arm, and may help us understand the extent of external validity of these results.

Balance, Attrition, and Implementation Checks In this section, we provide evidence to help assuage any concerns about the implementation of the experiment. First, we show balance in baseline covariates. Then, we present information on the attrition rates. Finally, we show teachers and managers have a strong understanding of the incentive schemes. Combined, this evidence suggests the experiment was implemented correctly.

Schools under the various performance evaluation treatments appear to be balanced along baseline covariates. Appendix table A.5 compares schools along numerous student and teacher baseline characteristics. Of 27 tests, one is statistically significant at the 10% level, and one is statistically significant at the 5% level, no more than we would expect by random chance. Results control for these few unbalanced variables.

Administrative data is available for all teachers and students who stay employed or enrolled during the year of the intervention. During this time, 23% of teachers leave the school system, which is very similar to the historical turnover rate. 88% of employed teachers completed the endline survey. While teachers were frequently reminded and encouraged to complete the survey, some chose not to. We do not see differences in these rates by treatment.

Teachers appear to understand their treatment assignment. Six months after the end of the intervention, we asked teachers to explain the key features of their treatment assignment. 60% of teachers could identify the key features of their raise treatment. Finally, most teachers stated that they came to fully understand what was expected of them in their given treatment within four months of the beginning of the information campaign.

3.4 Results

In the following section, we will look at the effect of two changes to the performance evaluation process: delinking performance evaluations from financial compensation for employees and increasing the time managers spend observing workers. For each, we will show the effects of the intervention on the extent of gender bias in evaluation scores. Finally,

we will see if these effects vary by manager characteristic, such as gender, age and extent of bias (as measured by survey questions).

Evidence of Financial Discrimination

Our main specification is:

$$\begin{aligned} PredictedRaise_{is} = \beta_0 + \beta_1 FinancialTreatment_s + \beta_2 Female_i \\ + \beta_3 FinancialTreatment_s * Female_i + \epsilon_{is} \end{aligned} \tag{3.9}$$

where the dependent variable is the predicted raise based on the employee's evaluation score. In the control group, this is the raise the employee would have received had the raise been based on the manager's evaluation and in the treatment this is the actual raise the teacher ended up receiving. $FinancialTreatment_s$ is a dummy for whether the employee's school, s, had financial stakes tied to the raise, and $Female$ is a dummy for whether the employee is female. The coefficients of interest is β_3 which tests whether men and women's scores are differentially affected by tying them to financial stakes. Standard errors are clustered at the school level (the unit of randomization).

Table 3.4 column (2) presents the results of eq. 3.9. First, we see that female teachers whose evaluation score did not actually determine their raise would have received, on average, a slightly lower raise than male teachers ($22.50 less or 6% lower) though the difference is not statistically significant. However, when the evaluation score does have a financial stake, we see that women receive significantly lower scores than men. In total female teachers' raises are $52.50 (14%) lower than male teachers' raises in the financial treatment. This suggests that when evaluation scores are tied to the employee's raise, this differentially impacts women.

Evidence of Statistical Discrimination

For the second treatment arm, we will first show that having managers do more classroom observations does actually improve the information they have about teacher quality. Table 3.5 shows the relationship between managers' beliefs about different aspects of teacher performance and their actual performance. Column (1) pools across all four aspects of teacher quality (attendance, disciplinary management of students, focus on analysis/inquiry skills and value-added) and columns (2) -(5) presents each of these components separately. We can see that, on average, managers seem to have fairly accurate information about teacher attendance and disciplinary management but are less accurate about the other aspects of teacher performance. Finally, column (6) shows the interaction between whether the teacher was assigned to be observed more frequently. We find that managers are much more accurate in evaluating teacher performance when they were required to observe them more frequently. This suggests that the treatment worked in improving the accuracy of information managers have about their employees.

Our key question then is when managers have more accurate information about employees, does this lower the extent of gender bias? Our main specification is:

$$PredictedRaise_{is} = \beta_0 + \beta_1 ObservationTreatment_i + \beta_2 Female_i + \beta_3 ObservationTreatment_i * Female_i + \epsilon_{is} \quad (3.10)$$

where the dependent variable is the predicted raise based on the employee's evaluation score, r_i, for employee, i. $ObservationTreatment_s$ is a dummy for whether the employee was assigned to be observed more frequently by their manager, and $Female$ is a dummy for whether the employee is female. The coefficient of interest is β_3 which tests whether men and women's scores are differentially affected by tying them to financial stakes. Standard errors are at the employee level (the unit of randomization).

We find that the observation treatment significantly reduces the extent of gender bias. Female teachers who were not observed receive a \$61 lower raise than male teachers, but that difference is cut in half when teachers were part of the observation treatment. Table 3.4 column (3) presents the results of eq. 3.10. It appears that introducing the observation treatment helps to mitigate some of the effects of the financial stakes, and when there are no financial stakes the information treatment does not have much of an effect because there is already no gender discrimination. Table 3.4 column (4) combines both the financial stakes and observation treatment.

3.5 Heterogeneity by Managers

We find that when evaluation scores are not tied to financial rewards for employees and managers have increased exposure to employees, we see no gap in the scores of male and female teachers. However, we might expect that the role information and financial stakes play in performance evaluations may vary by manager characteristic. To test this, we look at heterogeneous treatment effects by a variety of manager characteristics: gender, age, and extent of gender bias. We measure gender bias based on the managers' response to three questions. Managers rate how much they agree or disagree with the following statements:

- *Men are better suited than women to teach math and science*
- *When jobs are scarce, men should have more right to a job than women*
- *In general, it is better for a family if a woman has the main responsibility for taking care of the home and children rather than a man.*

Overall, we do not find a dramatic difference in the treatment effect by manager characteristic, though our standard errors are large, so we cannot reject relatively large effects in either direction. Table 3.6 and table 3.7 present results from eq. 3.9 and eq. 3.10 adding in an interaction with the respective manager covariate. One suggestive pattern we see is that the financial stakes actually have less of a negative effect on female ratings for managers who are more "biased" as measured from our survey. This suggests that even

managers who are not outwardly admitting to gender bias are still changing their evaluations when there are bigger consequences for those ratings. We do not find a differential response to the financial stakes by manager gender or age. We also do not find a differential response to the observation treatment manager gender, age or bias as measured by survey response.

3.6 Conclusion

This paper shows that even in settings with low levels of stated gender bias, we can have disparate wage outcomes for employees. We show that when employees' performance evaluations are tied to their end-of-year raise, female employees receive systematically lower evaluation scores. However, managers show less gender bias the more time they spend actually observing the employee, suggesting that better information can help correct gender bias. These effects are consistent across male and female managers.

3.7 Tables

Table 3.1: World Values Survey Summary Statistics

	Study Sample	World Values Survey Sample	
		South Asia	OECD
	(1)	(2)	(3)
When jobs are scarce, men should have more right to a job than women	4.1	1.7	3.5
On the whole, men make better business executives than women do	3.8	1.9	3.1
When a mother works for pay, the children suffer	3.7	1.8	2.7

Notes: This table presents the response to World Values Survey questions related to women in the workplace for our sample versus a representative sample. Responses vary from (1) strongly agree to (5) strongly disagree. A low score on each item then is characteristic of more gender bias. Column (1) is the average response from our study managers. Column (2) is the average for respondents in the World Value Survey for all South Asian countries. Column (3) is the average response across all OECD countries.

Table 3.2: Manager Rating by Vignette Characteristic

	Manager's rating (percentile)		
	(1)	(2)	(3)
Female name	-0.458	0.301	1.304
	(1.010)	(0.761)	(2.861)
Value-Added percentile		0.321***	0.323***
		(0.0177)	(0.0225)
Behavioral management percentile		0.164***	0.167***
		(0.0131)	(0.0275)
Attendance percentile		0.146***	0.151***
		(0.0148)	(0.0239)
Value added percentile*Female name			-0.00412
			(0.0246)
Behavioral management percentile*Female name			-0.00521
			(0.0392)
Attendance percentile*Female name			-0.0109
			(0.0348)
Constant	60.09***	28.49***	27.98***
	(1.150)	(1.918)	(2.631)
Observations	567	567	567
Dep. Var. Mean	59.86	59.86	59.86
Dep. Var. SD	18.13	18.13	18.13

Notes: This table shows the relationship between different vignette characteristics and the evaluation score managers gave them in the endline survey. During the endline survey managers are randomly provided vignettes of teachers to rate. The vignettes vary the gender and described productivity of the teacher along several dimensions. The outcome is the manager's rating of the teacher described in the vignette in percentile (ranging from 0-100). *Female name* is a dummy for whether the teacher in the vignette had a traditionally female name. *Value-added, behavioral management and attendance percentile* are the percentile the teacher in the vignette was in for each area of teacher performance. The possible values for these variables are 10, 50 and 90. Column (1) just includes the female name dummy. Column (2) controls for the other performance characteristics, and column (3) adds in an interaction between the gender of the name and the performance characteristics. Standard errors are clustered at the manager level (the unit of randomization). $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table 3.3: Manager Rating by Vignette and Manager Characteristic

	Manager's rating (percentile)					
	(1) Male	(2) Age	(3) Avg. Bias	(4) Math	(5) Jobs	(6) Family
Female name	0.556	1.773	5.022*	3.176	0.856	4.323**
	(0.846)	(4.942)	(2.843)	(2.253)	(1.996)	(2.173)
Interaction	0.718	-0.00753	2.041	0.693	0.904	1.565
	(3.335)	(0.137)	(1.688)	(1.284)	(1.425)	(1.086)
Interaction*Female name	-3.071	-0.0308	-2.168*	-1.292	-0.279	-1.680**
	(3.054)	(0.105)	(1.242)	(0.979)	(0.935)	(0.800)
Value-Added percentile	0.318***	0.318***	0.319***	0.319***	0.319***	0.317***
	(0.0187)	(0.0189)	(0.0187)	(0.0183)	(0.0188)	(0.0189)
Behavioral management percentile	0.168***	0.168***	0.168***	0.168***	0.168***	0.168***
	(0.0138)	(0.0140)	(0.0138)	(0.0138)	(0.0140)	(0.0137)
Attendance percentile	0.145***	0.145***	0.146***	0.146***	0.145***	0.146***
	(0.0158)	(0.0156)	(0.0157)	(0.0158)	(0.0158)	(0.0156)
Constant	28.49***	28.88***	24.07***	26.96***	26.79***	24.84***
	(2.057)	(7.032)	(3.966)	(3.080)	(3.158)	(3.411)
Observations	522	522	522	522	522	522
Dep. Var. Mean	59.86	59.86	59.86	59.86	59.86	59.86
Dep. Var. SD	18.13	18.13	18.13	18.13	18.13	18.13

Notes: This table shows the relationship between different vignette characteristics, the evaluation score managers gave them in the endline survey and the characteristic of the manager themselves. During the endline survey managers are randomly provided vignettes of teachers to rate. The vignettes vary the gender and described productivity of the teacher along several dimensions. The outcome is the manager's rating of the teacher described in the vignette in percentile (ranging from 0-100). *Female name* is a dummy for whether the teacher in the vignette had a traditionally female name. *Value-added, behavioral management and attendance percentile* are the percentile the teacher in the vignette was in for each area of teacher performance. The possible values for these variables are 10, 50 and 90. *Interaction* is a manager characteristic, with each column using a different characteristic. In column (1), *Interaction* is a dummy for if the manager is male. In column (2), it is the manager's age in years. In column (3), it is the manager's average score on the World Values survey gender bias questions, ranging from 1 (least gender biased) to 5 (most gender biased). In columns (4)-(6), the interaction is the manager's response to each individual question for the world values survey. Standard errors are clustered at the manager level (the unit of randomization). $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table 3.4: Raise Amount by Treatment and Gender

	Predicted Raise Amount (USD)			
	(1)	(2)	(3)	(4)
Female	-46.16***	-22.50	-60.84***	-26.44
	(6.783)	(13.92)	(13.82)	(21.05)
Financial Treatment		10.94		50.36
		(26.08)		(36.64)
Financial Treatment*Female		-30.49*		-66.53**
		(17.34)		(29.08)
Observation Treatment			-27.82	0.921
			(17.04)	(28.91)
Observation Treatment*Female			30.91*	9.674
			(18.34)	(31.01)
Financial Treatment*Observation Treatment				-55.69
				(37.71)
Financial Treatment*Observation Treatment*Female				50.01
				(41.35)
Observations	5051	4300	2626	2326
Clusters	.	263	.	158
Dep. Var. Mean	365.4	365.4	365.4	365.4
Dep. Var. SD	164.7	164.7	164.7	164.7

Notes: This table presents the relationship between the employee's performance evaluation score, the treatment status and gender. The dependent variable is the employee's evaluation score converted into the associated raise value in USD for that score. *Female* is a dummy for whether the employee is female. *Financial Treatment* is a dummy which is 1 if the teacher's school was assigned to have their evaluation determine their raise or 0 if their evaluation was just for feedback purposes. textitObservation Treatment is a dummy which is 1 if the teacher was randomly assigned be observed more frequently by their manager and 0 otherwise. Standard errors are clustered at the school level (the unit of randomization). $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table 3.5: Manager Beliefs by Treatment

	Manager Belief (z-score)					
	(1) All	(2) Attendance	(3) Discipline	(4) Analysis/Inquiry	(5) VA	(6) All
Teacher Outcome (z-score)	0.168***	0.192***	0.231**	0.136	-0.0435	0.0580
	(0.0433)	(0.0503)	(0.104)	(0.125)	(0.0831)	(0.0680)
Observation treatment						-0.0433
						(0.0900)
Teacher Outcome*Observation treatment						0.195*
						(0.1000)
Dep. Var. Mean	-0.0351	-0.0978	0.00316	0.0132	-0.0152	-0.0351
Dep. Var. SD	1.003	1.029	0.996	0.983	0.988	1.003
Observations	702	250	143	143	166	594
Grade Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table presents the relationship between teacher outcomes and principals beliefs about those outcomes. There are four outcomes principals rate teachers on: attendance, management of student discipline, incorporation of analysis and inquiry skills and value-added. *Principal beliefs* are from principal endline survey data. Actual teacher outcomes come from administrative and classroom observation data. Attendance is measured using biometric clock in and out data. Discipline and analysis/inquiry are rates via classroom observations. *Observation Treatment* is a dummy which is 1 if the teacher was randomly assigned be observed more frequently by their manager and 0 otherwise. Column (1) pools all four outcomes. Column (2)-(5) separates the results by outcome type. Column (6) pools across all four outcomes and add in an interaction with treatment status. Standard errors are clustered at the manager level. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table 3.6: Effect of Financial Stakes by Manager Type

	Predicted Raise Amount (USD)					
	(1) Male	(2) Age	(3) Avg. Bias	(4) Math	(5) Jobs	(6) Family
Female	-31.77*	-88.23	-44.40	-55.51	-38.07	-18.46
	(16.26)	(103.0)	(66.30)	(66.46)	(48.12)	(36.18)
Interaction	-84.65	-3.522*	-17.54	-15.80	-9.782	-4.119
	(64.42)	(1.967)	(43.21)	(41.70)	(25.62)	(29.20)
Financial Treatment	22.93	-141.6	197.5	126.6	158.8	116.2
	(39.60)	(198.7)	(149.3)	(125.2)	(109.2)	(120.2)
Financial Treatment*Female	-61.62**	19.73	-201.7**	-131.6	-194.3**	-136.1*
	(30.06)	(164.7)	(99.26)	(93.66)	(85.51)	(74.71)
Interaction*Financial Treatment	56.39	3.678	-76.17	-39.97	-67.75	-36.53
	(76.74)	(3.890)	(59.53)	(51.85)	(45.28)	(39.77)
Interaction*Female	47.03	1.355	8.217	14.63	6.094	-3.274
	(49.76)	(2.082)	(27.58)	(29.44)	(20.89)	(13.30)
Interaction*Financial Treatment*Female	-1.904	-1.824	63.31	27.27	68.35*	30.47
	(64.59)	(3.277)	(40.08)	(37.85)	(36.45)	(25.17)
Constant	415.7***	571.2***	444.9***	438.5***	425.6***	417.4***
	(25.15)	(103.2)	(107.1)	(91.62)	(63.23)	(86.96)
Observations	3650	3650	3650	3650	3650	3650
Clusters	208	208	208	208	208	208
Dep. Var. Mean	368.4	368.4	368.4	368.4	368.4	368.4
Dep. Var. SD	176.3	176.3	176.3	176.3	176.3	176.3

Notes: This table presents the relationship between the employee's performance evaluation score, the treatment status and manager characteristics. The dependent variable is the employee's evaluation score converted into the associated raise value in USD for that score. *Female* is a dummy for whether the employee is female. *Financial Treatment* is a dummy which is 1 if the teacher's school was assigned to have their evaluation determine their raise or 0 if their evaluation was just for feedback purposes. *Interaction* is a manager characteristic, with each column using a different characteristic. In column (1), *Interaction* is a dummy for if the manager is male. In column (2), it is the manager's age in years. In column (3), it is the manager's average score on the World Values survey gender bias questions, ranging from 1 (least gender biased) to 5 (most gender biased). In columns (4)-(6), the interaction is the manager's response to each individual question for the world values survey. Standard errors are clustered at the school level (the unit of randomization). $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table 3.7: Effect of Information by Manager Type

	Predicted Raise Amount (USD)					
	(1) Male	(2) Age	(3) Avg. Bias	(4) Math	(5) Jobs	(6) Family
Female	-60.06***	-225.0	-91.40	-100.3*	-117.7**	-30.64
	(19.47)	(191.0)	(58.34)	(54.16)	(46.15)	(39.70)
Interaction	-99.09*	-5.532	-36.96	-32.40	-25.69	-9.613
	(54.62)	(4.113)	(29.40)	(29.00)	(21.10)	(19.39)
Observation Treatment	-12.47	-212.9	22.77	4.496	-31.04	30.11
	(27.35)	(194.5)	(87.14)	(65.33)	(60.49)	(73.69)
Observation Treatment*Female	18.25	266.0	-21.73	16.57	60.82	-53.11
	(29.29)	(228.2)	(102.1)	(68.18)	(68.37)	(83.01)
Interaction*Observation Treatment	48.14	4.318	-14.03	-7.731	10.41	-13.37
	(103.5)	(4.062)	(44.75)	(29.80)	(29.92)	(33.83)
Interaction*Female	80.48	3.609	16.45	21.53	31.23	-9.175
	(52.47)	(3.952)	(25.28)	(23.73)	(19.48)	(16.29)
Interaction*Observation Treatment*Female	-40.40	-5.288	16.61	0.621	-22.24	25.05
	(112.3)	(4.670)	(47.72)	(30.75)	(32.75)	(34.35)
Constant	418.2***	674.0***	494.1***	480.7***	464.0***	435.7***
	(22.38)	(198.4)	(65.65)	(64.46)	(50.80)	(47.43)
Observations	2614	2614	2614	2614	2614	2614
Clusters	147	147	147	147	147	147
Dep. Var. Mean	368.4	368.4	368.4	368.4	368.4	368.4
Dep. Var. SD	176.3	176.3	176.3	176.3	176.3	176.3

Notes: This table presents the relationship between the employee's performance evaluation score, the treatment status and manager characteristics. The dependent variable is the employee's evaluation score converted into the associated raise value in USD for that score. *Female* is a dummy for whether the employee is female. *Observation Treatment* is a dummy which is 1 if the teacher was randomly assigned be observed more frequently by their manager and 0 otherwise. *Interaction* is a manager characteristic, with each column using a different characteristic. In column (1), *Interaction* is a dummy for if the manager is male. In column (2), it is the manager's age in years. In column (3), it is the manager's average score on the World Values survey gender bias questions, ranging from 1 (least gender biased) to 5 (most gender biased). In columns (4)-(6), the interaction is the manager's response to each individual question for the world values survey. Standard errors are at the teacher level (the unit of randomization). $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Bibliography

George A. Akerlof. The Market for "Lemons": Quality Uncertainty and the Market Mechanism. The Quarterly Journal of Economics, 84(3):488–500, August 1970.

Tahir Andrabi and Christina Brown. Subjective and Objective Incentives and Employee Productivity. Working Paper, page 52, July 2020.

ASER. Annual Status of Education Report Pakistan. 2019. URL aserpakistan.org.

Nava Ashraf, Oriana Bandiera, Edward Davenport, and Scott S. Lee. Losing Prosociality in the Quest for Talent? Sorting, Selection, and Productivity in the Delivery of Public Services. American Economic Review, 110(5):1355–1394, May 2020. ISSN 0002-8282. doi: 10.1257/aer.20180326. URL https://pubs.aeaweb.org/doi/10.1257/aer.20180326.

Mehtabul Azam and Geeta Kingdon. Assessing Teacher Quality in India. Working Paper, page 31, October 2014. URL AvailableatSSRN:https://ssrn.com/abstract=2512933.

George Baker. Distortion and Risk in Optimal Incentive Contracts. The Journal of Human Resources, 37(4):728, 2002. ISSN 0022166X. doi: 10.2307/3069615. URL https://www.jstor.org/stable/3069615?origin=crossref.

Gadi Barlevy and Derek Neal. Pay for Percentile. American Economic Review, 102(5): 1805–1831, August 2012. ISSN 0002-8282. doi: 10.1257/aer.102.5.1805. URL https://pubs.aeaweb.org/doi/10.1257/aer.102.5.1805.

Natalie Bau and Jishnu Das. Teacher Value Added in a Low-Income Country. American Economic Journal: Economic Policy, 12(1):62–96, February 2020. ISSN 1945-7731, 1945-774X. doi: 10.1257/pol.20170243. URL https://pubs.aeaweb.org/doi/10.1257/pol.20170243.

Barbara Biasi. Unions, Salaries, and the Market for Teachers: Evidence from Wisconsin. SSRN Electronic Journal, 2017. ISSN 1556-5068. doi: 10.2139/ssrn.2942134. URL http://www.ssrn.com/abstract=2942134.

Nicholas Bloom, Renata Lemos, Raffaella Sadun, and John Van Reenen. Does Management Matter in schools? The Economic Journal, 125(584):647–674, 2015. doi: 10.1111/ecoj.12267. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ecoj.12267. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecoj.12267.

Christina Brown and Tahir Andrabi. Overcoming Adverse Selection through Performance Pay. Working Paper, July 2020.

Miriam Bruhn and David McKenzie. In Pursuit of Balance: Randomization in Practice in Development Field Experiments. American Economic Journal: Applied Economics, 1(4): 34, October 2009.

David L Burrell. Relationships Among Teachers' Efficacy, Teachers' Locus-of-control, and Student Achievement, 1994. URL https://dc.etsu.edu/etd/2646.

Raj Chetty, John N Friedman, and Jonah E Rockoff. Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. American Economic Review, 104(9): 2593–2632, September 2014a.

Raj Chetty, John N. Friedman, and Jonah E. Rockoff. Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. American Economic Review, 104(9):2633–2679, September 2014b. ISSN 0002-8282. doi: 10.1257/aer.104.9.2633. URL https://pubs.aeaweb.org/doi/10.1257/aer.104.9.2633.

Axel Engellandt and Regina Riphahn. Evidence on Incentive Effects of Subjective Performance Evaluations. Industrial & Labor Relations Review, 64, 2011. doi: 10.2307/41149491.

Roland G. Fryer. Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. Journal of Labor Economics, 31(2):373–407, April 2013. ISSN 0734-306X, 1537-5307. doi: 10.1086/667757. URL https://www.journals.uchicago.edu/doi/10.1086/667757.

Michael Gibbs, Kenneth A. Merchant, Wim A. Van der Stede, and Mark E. Vargus. Determinants and Effects of Subjectivity in Incentives. The Accounting Review, 79(2): 409–436, 2004. URL http://www.jstor.org/stable/3203250.

Sarena F. Goodman and Lesley J. Turner. The Design of Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program. Journal of Labor Economics, 31(2):409–420, April 2013. ISSN 0734-306X, 1537-5307. doi: 10.1086/668676. URL https://www.journals.uchicago.edu/doi/10.1086/668676.

Bruce C. Greenwald. Adverse Selection in the Labour Market. The Review of Economic Studies, 53(3):325, July 1986. ISSN 00346527. doi: 10.2307/2297632. URL https://academic.oup.com/restud/article-lookup/doi/10.2307/2297632.

The World Bank Group. World Development Report 2018: Learning to Realize Education's Promise. World Bank, Washington, DC, 2018. ISBN 978-1-4648-1096-1.

C. Kirabo Jackson. What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes. Journal of Political Economy, 126(5):2072–2107, October

2018. ISSN 0022-3808. doi: 10.1086/699018. URL https://doi.org/10.1086/699018. Publisher: The University of Chicago Press.

Brian A. Jacob and Lars Lefgren. Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. Journal of Labor Economics, 26 (1):101–136, January 2008. ISSN 0734-306X, 1537-5307. doi: 10.1086/522974. URL https://www.journals.uchicago.edu/doi/10.1086/522974.

Andrew C. Johnston. Teacher Preferences, Working Conditions, and Compensation Structure. SSRN Electronic Journal, 2020. ISSN 1556-5068. doi: 10.2139/ssrn.3532779. URL https://www.ssrn.com/abstract=3532779.

Thomas Kane and Douglas Staiger. Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. Technical Report w14607, National Bureau of Economic Research, Cambridge, MA, December 2008. URL http://www.nber.org/papers/w14607.pdf.

Adnan Q. Khan, Asim Ijaz Khwaja, and Benjamin A. Olken. Making Moves Matter: Experimental Evidence on Incentivizing Bureaucrats through Performance-Based Postings. American Economic Review, 109(1):237–270, January 2019. ISSN 0002-8282. doi: 10.1257/aer.20180277. URL https://pubs.aeaweb.org/doi/10.1257/aer.20180277.

Victor Lavy. Using Performance-Based Pay to Improve the Quality of Teachers. The Future of Children, 17(1,):87–109, 2007. URL http://www.jstor.org/stable/4150021.

Victor Lavy. Performance Pay and Teachers' Effort, Productivity, and Grading Ethics. American Economic Review, 99(5):1979–2011, December 2009. ISSN 0002-8282. doi: 10.1257/aer.99.5.1979. URL https://pubs.aeaweb.org/doi/10.1257/aer.99.5.1979.

Edward P Lazear. Performance Pay and Productivity. The American Economic Review, 90 (5):66, 2000.

Edward P. Lazear and Robert L. Moore. Incentives, Productivity, and Labor Contracts. The Quarterly Journal of Economics, 99(2):23, May 1984.

Edward P Lazear and Paul Oyer. Personnel Economics. In The Handbook of Organizational Economics, pages pp. 479–519. Princeton University Press., Princeton; Oxford, December 2012.

Clare Leaver, Owen Ozier, Pieter Serneels, and Andrew Zeitlin. Recruitment, effort, and retention effects of performance contracts for civil servants: Experimental evidence from Rwandan primary schools. RISE Programme Working Paper, page 41, June 2019.

A. Leigh. The Economics and Politics of Teacher Merit Pay. CESifo Economic Studies, 59(1):1–33, March 2013. ISSN 1610-241X, 1612-7501. doi: 10.1093/cesifo/ifs007. URL https://academic.oup.com/cesifo/article-lookup/doi/10.1093/cesifo/ifs007.

Karthik Muralidharan and Venkatesh Sundararaman. Teacher Performance Pay: Experimental Evidence from India. Journal of Political Economy, 119(1):39–77, February 2011. ISSN 0022-3808, 1537-534X. doi: 10.1086/659655. URL https://www.journals.uchicago.edu/doi/10.1086/659655.

National Center for Education Statistics. Schools and Staffing Survey, 2010-2011: [United States]. U.S. Dept. of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 2011. URL https://nces.ed.gov/surveys/sass/index.asp.

Paul Oyer and Scott Schaefer. Personnel Economics: Hiring and Incentives. volume 4B, pages 1769–1823. Elsevier, 1 edition, 2011. URL https://EconPapers.repec.org/RePEc:eee:labchp:5-20.

Lam D. Pham, Tuan D. Nguyen, and Matthew G. Springer. Teacher Merit Pay: A Meta-Analysis. American Educational Research Journal, 0(0):0002831220905580, February 2020. doi: 10.3102/0002831220905580. URL https://doi.org/10.3102/0002831220905580. _eprint: https://doi.org/10.3102/0002831220905580.

Robert C Pianta, Bridget K Hamre, and Susan Mintz. Classroom assessment scoring system: Secondary manual. Teachstone, 2012.

Michael J. Podgursky and Matthew G. Springer. Teacher performance pay: A review. Journal of Policy Analysis and Management, 26(4):909–950, 2007. ISSN 02768739, 15206688. doi: 10.1002/pam.20292. URL http://doi.wiley.com/10.1002/pam.20292.

Canice Prendergast. The Provision of Incentives in Firms. Journal of Economic Literature, 37(1):7–63, 1999. URL http://www.jstor.org/stable/2564725.

Canice Prendergast. The Motivation and Bias of Bureaucrats. The American Economic Review, 97(1):18, 2007.

Canice Prendergast and Robert Topel. Discretion and bias in performance evaluation. European Economic Review, 37(2-3):355–365, April 1993. ISSN 00142921. doi: 10.1016/0014-2921(93)90024-5. URL https://linkinghub.elsevier.com/retrieve/pii/0014292193900245.

Jonah E. Rockoff and Cecilia Speroni. Subjective and Objective Evaluations of Teacher Effectiveness. The American Economic Review, 100(2,):261–266, 2010. URL http://www.jstor.org/stable/27805001.

Evan K Rose, Jonathan Schellenberg, and Yotam Shem-Tov. The Effects of Teacher Quality on Criminal Behavior. Working Paper, page 63, May 2019.

Jesse Rothstein. Teacher Quality Policy When Supply Matters. American Economic Review, 105(1):100–130, January 2015. ISSN 0002-8282. doi: 10.1257/aer.20121242. URL https://pubs.aeaweb.org/doi/10.1257/aer.20121242.

A D Roy. Some Thoughts on the Distribution of Earnings. Oxford Economic Papers, 3(2): 135–146, June 1951. URL http://www.jstor.com/stable/2662082.

Matthew G Springer, Dale Ballou, Laura S Hamilton, Vi-Nhuan Le, J R Lockwood, Daniel F McCaffrey, Matthew Pepper, and Brian M Stecher. Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching. Technical report, National Center on Performance Incentives at Vanderbilt University, Nashville, TN, September 2010.

Douglas O Staiger and Jonah E Rockoff. Searching for Effective Teachers with Imperfect Information. Journal of Economic Perspectives, 24(3):97–118, August 2010. ISSN 0895-3309. doi: 10.1257/jep.24.3.97. URL http://pubs.aeaweb.org/doi/10.1257/jep.24.3.97.

The World Bank Group. Systems Approach for Better Education Results (SABER). 2018. URL http://saber.worldbank.org/.

The World Bank Group. Enterprise Surveys. 2019. URL http://www.enterprisesurveys.org.

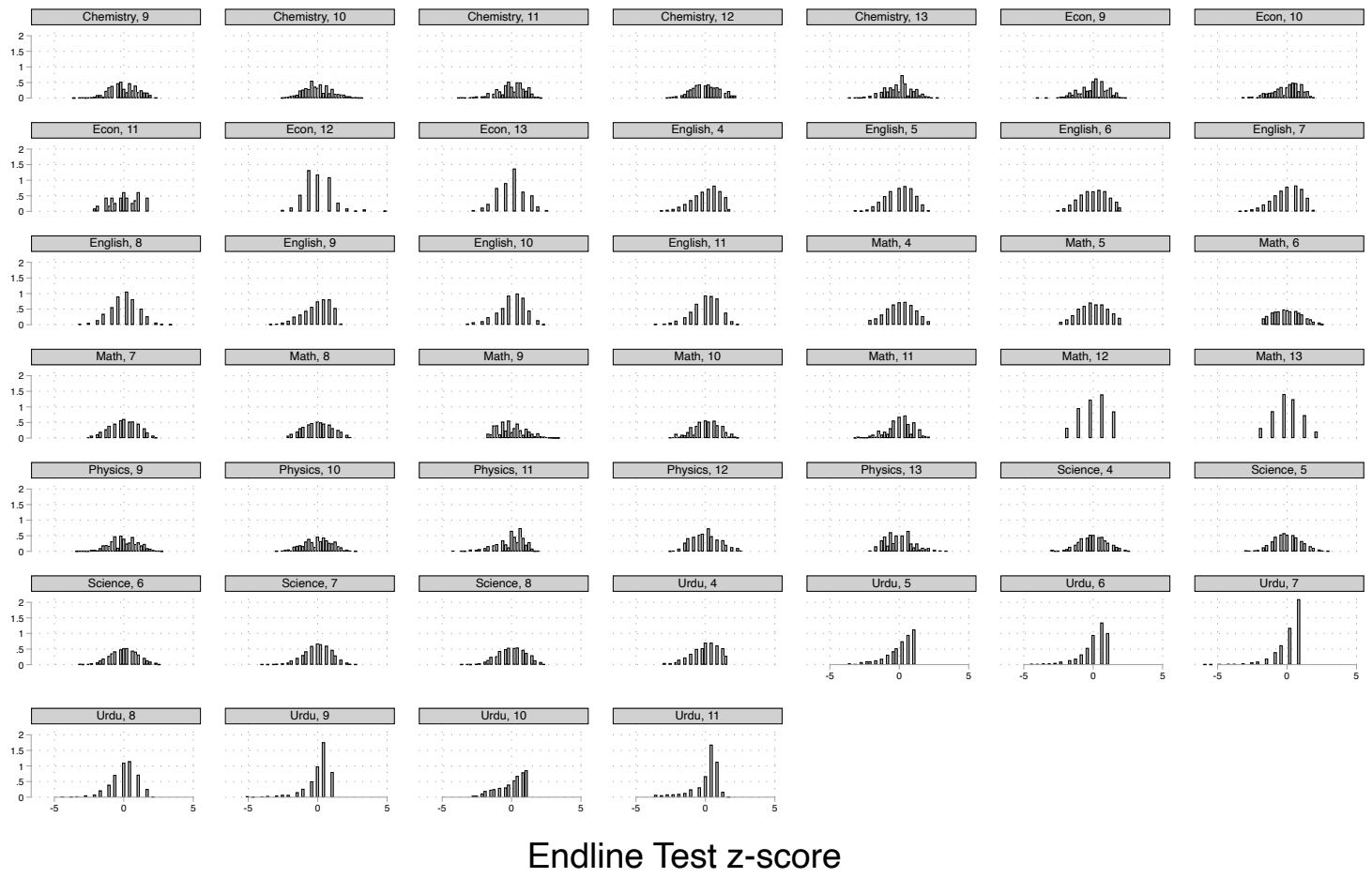
World Bank Group. Country Policy And Institutional Assessment Dataset. 2018. URL http://datatopics.worldbank.org/cpia/.

Appendix A

Appendix

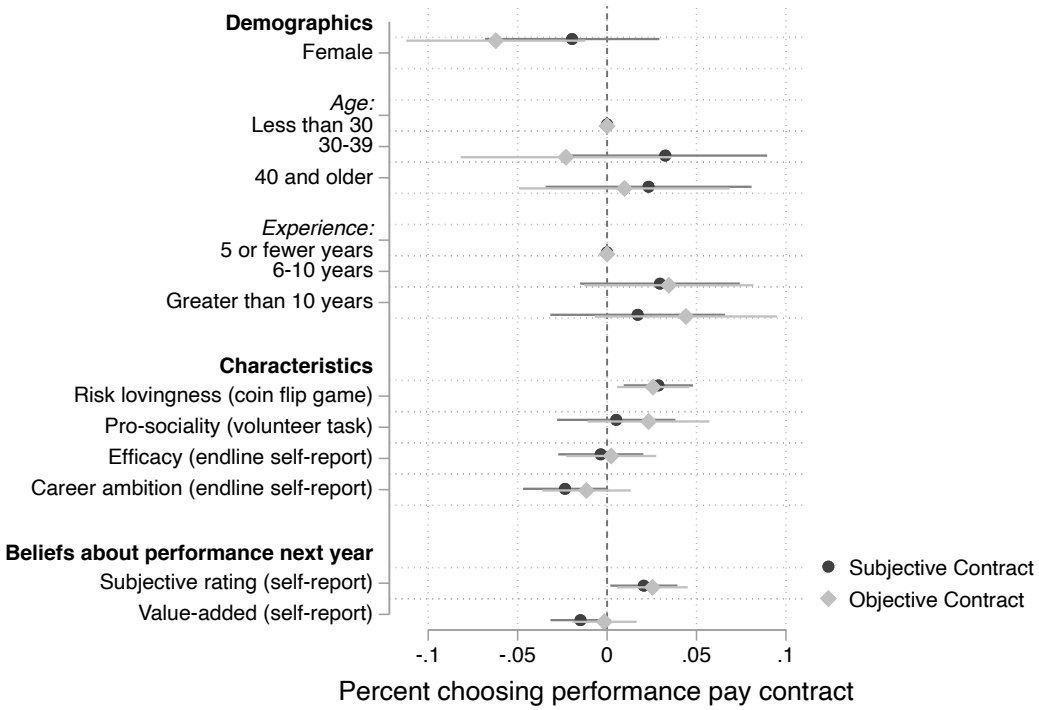
A.1 Supplementary Chap 1 Tables and Figures

Figure A.1: Distribution of Endline Test Scores



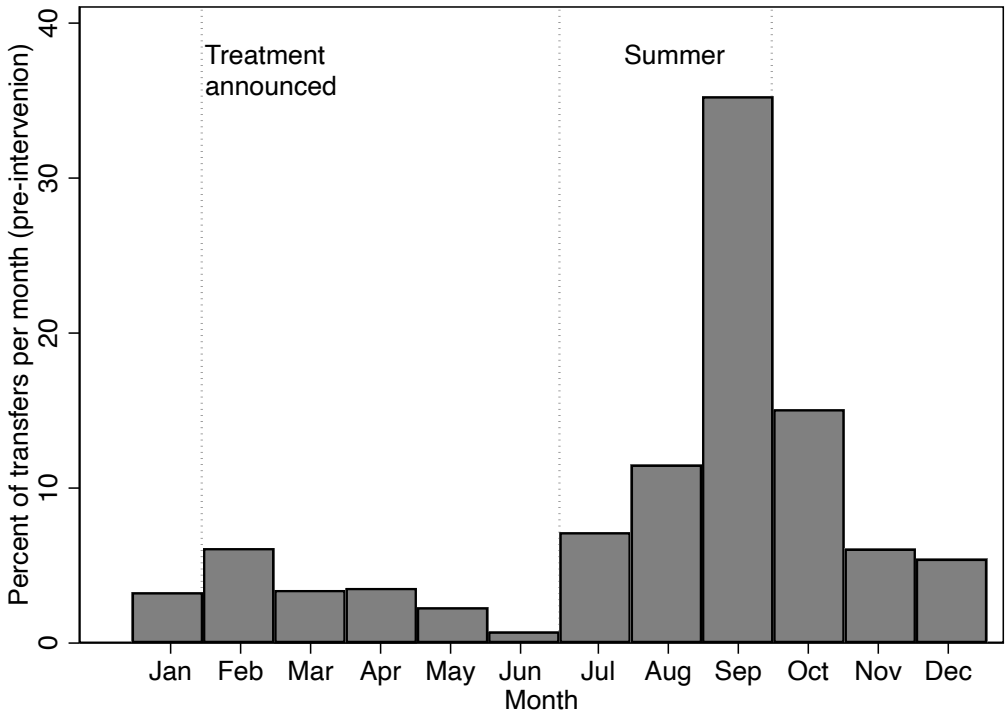
Notes: This figure presents the standardized distribution of student scores across each exam administered at endline. The endline test was conducted in January 2019 across grades 4-13 in English, Urdu, Math, Science and Economics. In grades 9-13, students took the science exam in the class they were currently enrolled, either Chemistry or Physics.

Figure A.2: Predictors of contract choice



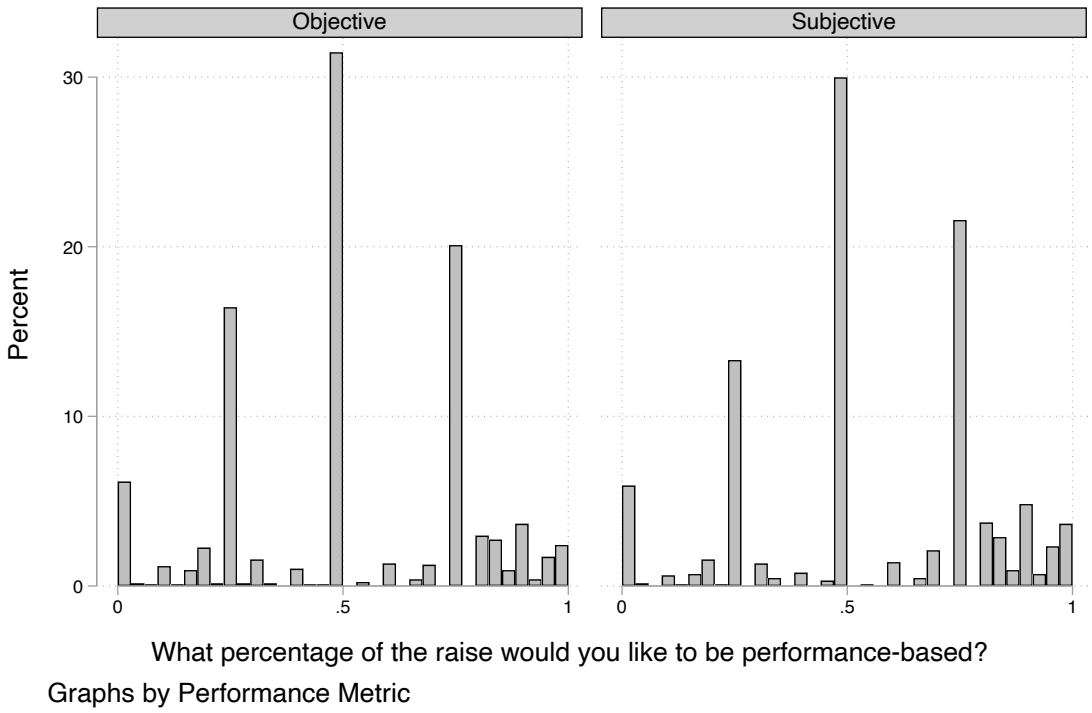
Notes: This figure presents coefficients of bi-variate regressions of teacher's contract choice on teacher demographics, characteristics and beliefs. Teacher's contract choice is a dummy for whether they selected a performance pay or flat pay contract. All independent variables other than gender, age and experience are standardized z-scores. Estimates in black are for the choice between subjective (principal evaluation based) performance pay versus flat pay (value-added based). Estimates in gray are for objective performance pay versus flat pay. Data is at the teacher-decision level, as teachers are asked to choose between performance and flat pay, first using an objective performance measure, then a subjective performance measure. Demographic data come from school administrative records. Characteristics (except efficacy and career ambition), beliefs and contract choice come from a baseline survey with 2,481 teachers.

Figure A.3: Teacher transfers across campuses within school system



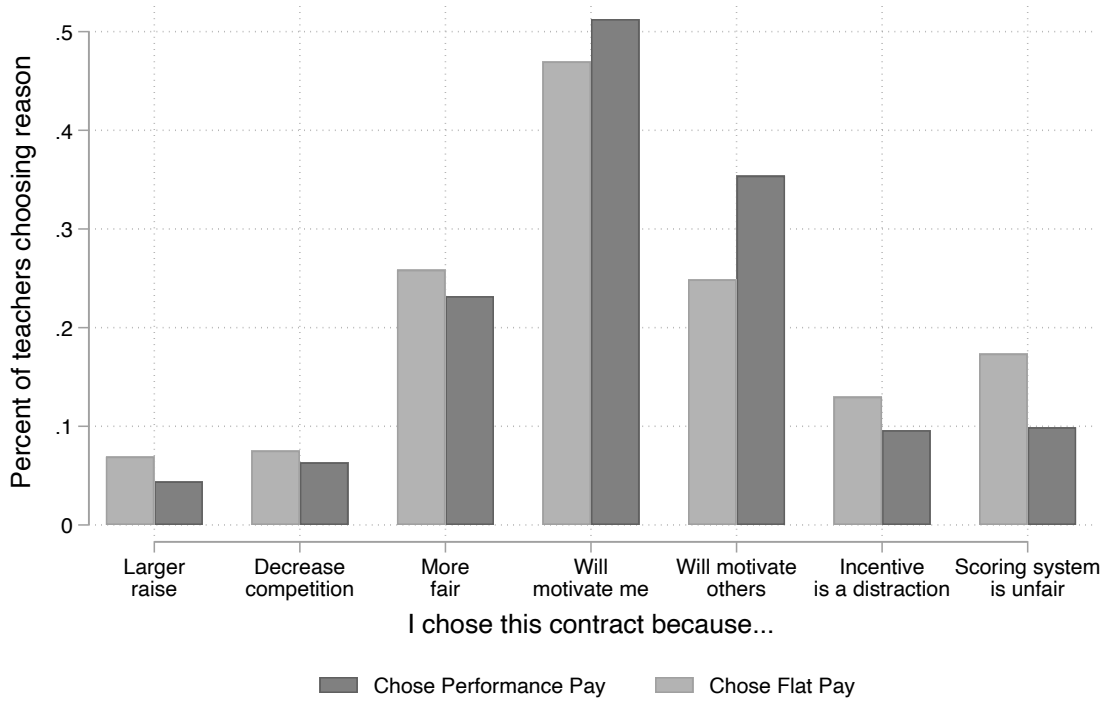
Notes: This figure plots the percent of transfers across schools within the system by month. Transfer data is from administrative schools records from 2015, prior to the intervention.

Figure A.4: Distribution of contract choice by performance metric



Notes: These figures plot teachers' survey response to the contract choice question. We ask teachers: *We can think of a raise as being a combination of two parts: the "flat" part that everyone gets regardless of their [subjective/objective] score and the "performance" part where those with higher [subjective/objective] scores receive more than those with low [subjective/objective] scores. What percentage of the raise would you like to be flat?"* The graphs plot 1 - the teacher's response. Data was collected during the baseline in October 2017.

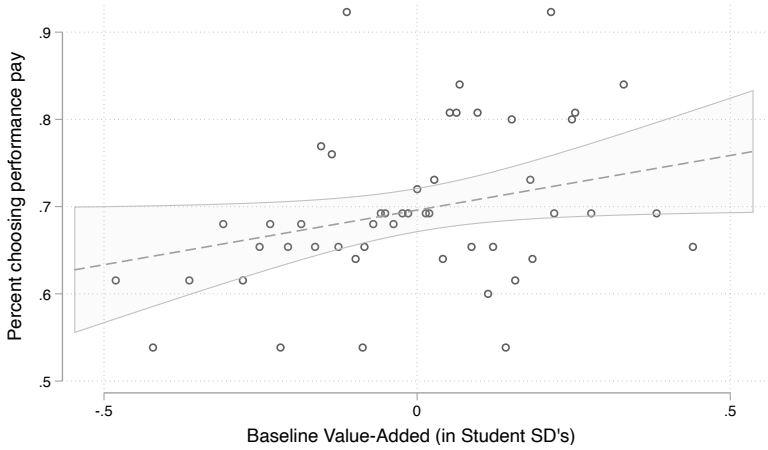
Figure A.5: Teachers stated reasons for selecting performance pay or flat pay contract



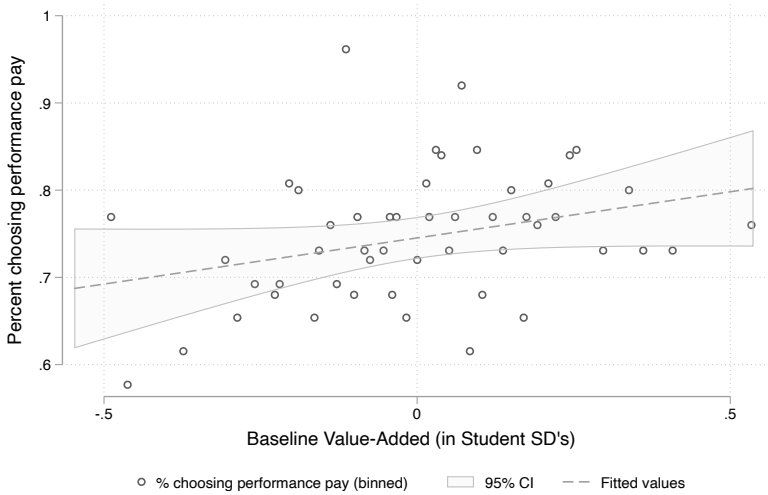
Notes: This figure plots teachers responses to the question *Why did you select this contract?*. The graph shows the percent of teachers that selected each reason. Teachers are allows to select multiple reasons, if applicable. The light gray bars plot responses for teachers who chose a flat pay contract. The dark gray bars plot responses for teachers who chose performance pay contracts.

Figure A.6: Relationship between Value-Added and Contract Choice

Panel A: Objective Performance Metric



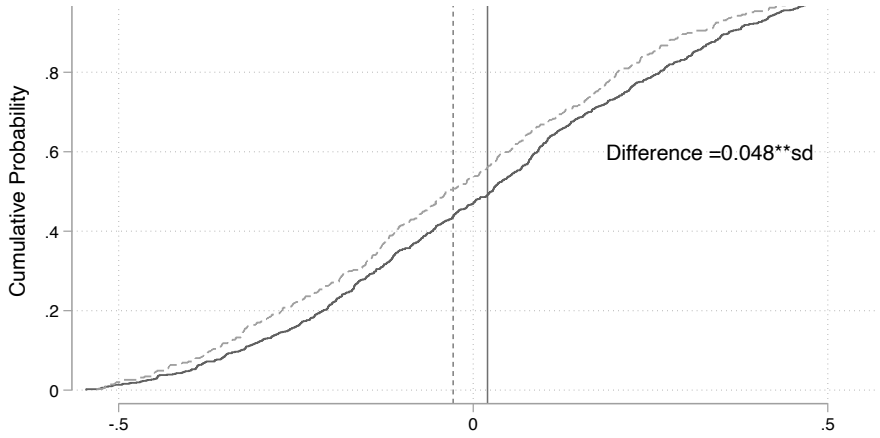
Panel B: Subjective Performance Metric



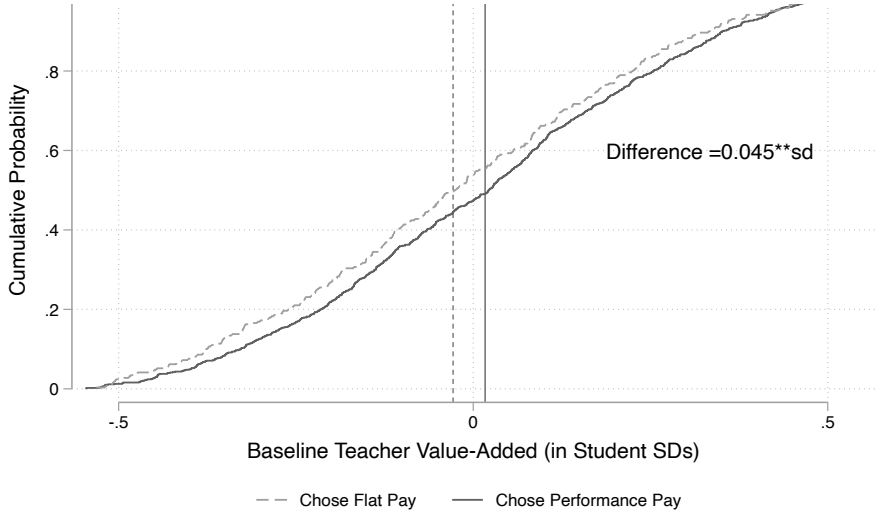
Notes: These figures plot the relationship between teacher quality as measured by baseline value-added and teachers contract choice. The graph plots binned values of *Teacher Baseline Value-Added* by the percent of teachers in that bin that chose performance pay. Panel A presents results for the choice between objective (value-added based) performance pay versus flat pay. Panel B presents results for the choice between subjective (principal evaluation based) performance pay versus flat pay. Choice data comes from the contract choice exercise conducted in October 2017. Value-added is calculated using two years of administrative data prior to the start of the intervention.

Figure A.7: Cumulative Distribution Function of Baseline Value-Added by Contract Choice

Panel A: Objective Performance Metric



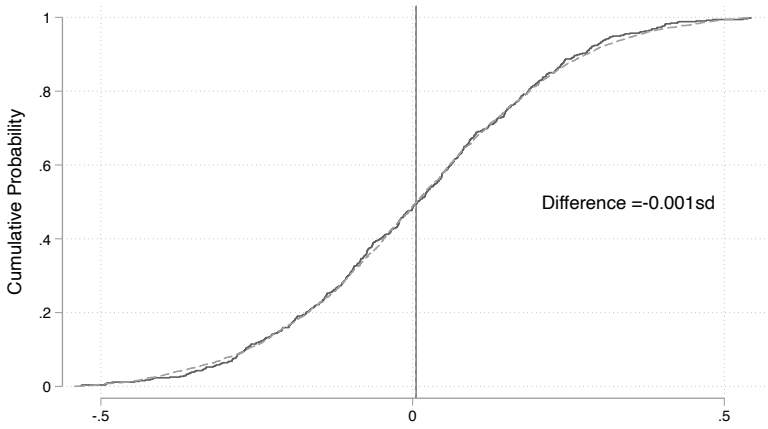
Panel B: Subjective Performance Metric



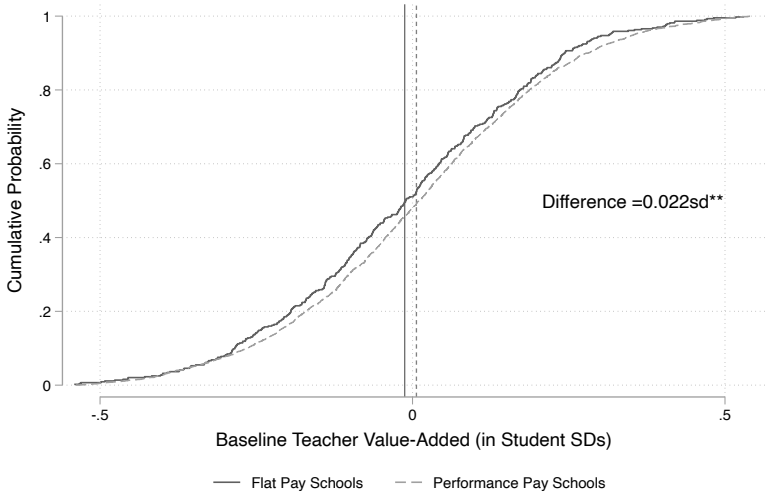
Notes: This figure plots the CDF of baseline teacher value-added for teachers who chose performance pay (solid line) versus flat pay (dotted line). Panel A presents results for the choice between objective (value-added based) performance pay versus flat pay. Panel B presents results for the choice between subjective (principal evaluation based) performance pay versus flat pay. Choice data comes from the contract choice exercise conducted in October 2017. Value-added is calculated using two years of administrative data prior to the start of the intervention. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Figure A.8: CDF of Teacher Baseline Value-Added by School Treatment and Year

Panel A: December 2017 (Baseline)

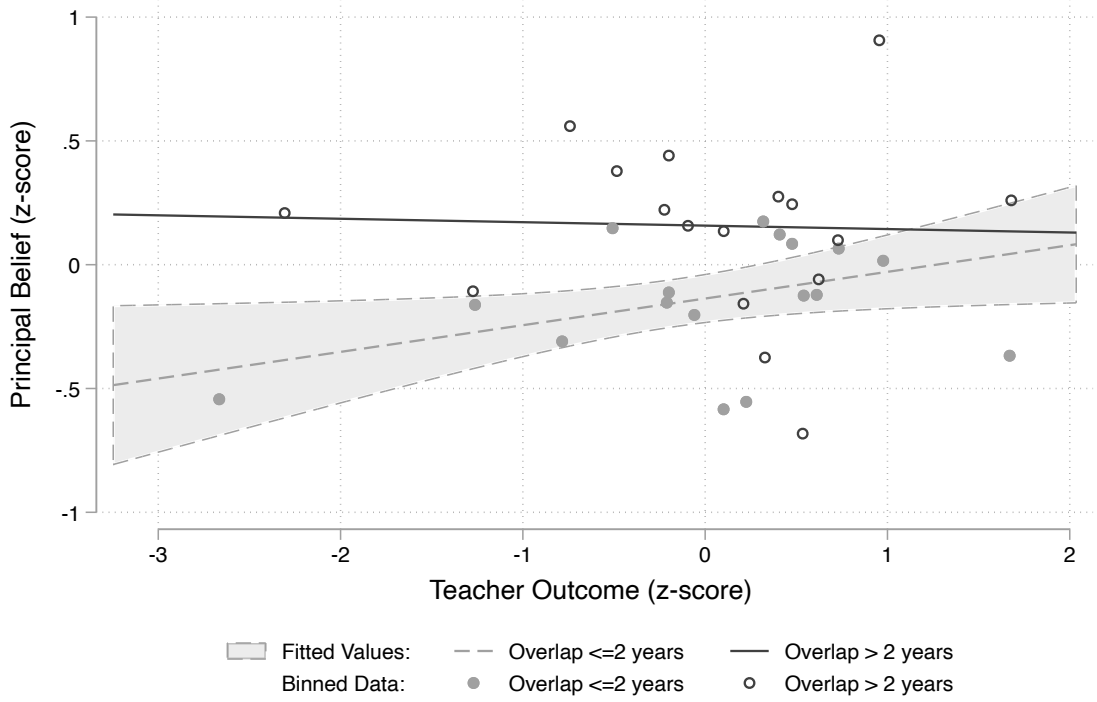


Panel B: December 2018 (One year after treatment announcement



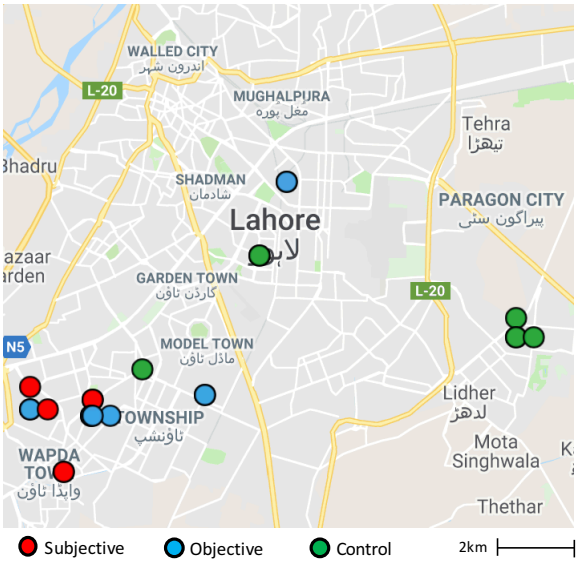
*Notes:*These figures plots the CDF of baseline teacher value-added for teachers in performance pay versus flat pay schools. Panel A provides the distribution in December 2017 (one month before the treatments are announced). Panel B provides the distribution in December 2018 (11 months after the treatments are announced). Teacher employment data comes from school administrative records. Value-added is calculated using two years of administrative data prior to the start of the intervention. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Figure A.9: Principal Beliefs about Teacher Outcome by Overlap of Principal and Teacher



Notes: This figure presents principals' beliefs about teacher quality versus their actual performance. *Principal beliefs* are measured in z-scores and come from endline surveys with principals. *Teacher outcome* is the the teacher's z-score in each of four outcomes: value-added, attendance, behavioral management and use of analysis/inquiry. Value-added is calculated using two years of administrative data prior to the start of the intervention. Standard errors are clustered at the school level. Attendance comes from bio-metric clock in and out data. The last two outcomes come from classroom video data. The results are split by whether the principal has worked at the same school with the teacher for two years or less (dotted line) or more than two years (solid line).

Figure A.10: Treatment Distribution Map, Lahore



Notes: The figures shows the location of treatment versus control performance pay assignments in one of the cities in our study.

Table A.1: Baseline Covariates

Variable	(1) Control N/ [Clusters]	Mean/ SE	(2) Objective Treatment N/ [Clusters]	Mean/ SE	(3) Subjective Treatment N/ [Clusters]	Mean/ SE	T-test Difference (1)-(2)	(1)-(3)	(2)-(3)
Panel A: Teacher Characteristics									
Performance evaluation score	656 [40]	3.360 (0.030)	384 [32]	3.362 (0.039)	3566 [139]	3.338 (0.010)	-0.002	0.022	0.024
Salary (USD)	920 [40]	5417.984 (313.504)	535 [32]	5125.462 (295.013)	4928 [145]	5329.416 (124.042)	292.523	88.569	-203.954
Age	921 [40]	36.591 (0.738)	539 [32]	36.083 (0.846)	4926 [145]	36.630 (0.298)	0.507	-0.039	-0.546
Years of experience	918 [40]	5.505 (0.277)	534 [32]	5.487 (0.425)	4897 [145]	5.725 (0.156)	0.019	-0.220	-0.238
Panel B: Student Test Scores									
Math Test Z-Score	9959 [40]	0.071 (0.070)	5292 [33]	-0.146 (0.065)	51775 [137]	-0.014 (0.026)	0.217**	0.085	-0.132*
Urdu Test Z-Score	9702 [40]	0.041 (0.072)	5259 [33]	-0.048 (0.063)	50915 [138]	-0.002 (0.028)	0.089	0.043	-0.046
English Test Z-Score	9755 [40]	0.017 (0.056)	5289 [33]	-0.049 (0.050)	51356 [137]	0.002 (0.032)	0.067	0.016	-0.051
Social Studies Test Z-Score	9171 [40]	0.041 (0.046)	5030 [33]	-0.064 (0.056)	49411 [137]	0.007 (0.022)	0.105	0.033	-0.071
Science Test Z-Score	9636 [40]	-0.010 (0.041)	5065 [33]	-0.064 (0.042)	50268 [137]	0.001 (0.024)	0.055	-0.011	-0.066

Notes: This table summarizes teacher and student characteristics before the experiment. The table reports mean values of each variable for each treatment group. The final three columns report mean differences between treatment group. Panel A presents teacher demographics as of September 2017. Panel B presents student test scores from yearly exams conducted in June 2017. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table A.2: Treatment Effect by Contract Choice

	Endline Test (z-score)				
	(1)	(2)	(3)	(4)	(5)
Assigned Perf Pay Treat	0.0660	-0.0163	-0.0171	0.0630	-0.0170
	(0.0408)	(0.0615)	(0.0617)	(0.0421)	(0.0643)
% Perf Pay		-0.0896	-0.0922		-0.0887
		(0.0678)	(0.0684)		(0.0663)
% Perf Pay* Assigned Perf Pay Treat		0.157**	0.159**		0.153*
		(0.0773)	(0.0774)		(0.0773)
Principal Rating of Teacher			0.00419		
			(0.0100)		
Baseline Value-Added				0.0282	0.0334
				(0.107)	(0.106)
Baseline Value-Added*Assigned Perf Pay Treat				-0.0729	-0.0844
				(0.129)	(0.127)
Control Mean	7.94e-10	7.94e-10	-0.00377	-0.00761	-0.00761
Control SD	1.000	1.000	0.999	0.997	0.997
Clusters	114	114	114	109	109
Observations	144009	144009	144009	126989	126989
Randomization Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes
Baseline	Yes	Yes	Yes	Yes	Yes

Notes: This table presents the treatment effect of performance pay contracts on endline test scores by teacher characteristics. The outcome is students' standardized z-score from the endline test conducted in January 2019. *Treated* is a dummy for whether a teacher taught at a school assigned to performance pay at baseline. *Chose Performance Pay* is a dummy variable for whether a teacher chose objective performance pay or flat pay during the baseline choice exercise. *Principal Rating of Teacher* is the baseline subjective rating z-score of the teacher by their principal. Column (1) presents the treatment effect for all teachers. Column (2) and (4) presents heterogeneity in treatment effect by contract choice and value-added, respectively. Column (5) combines the two and column (3) controls for principal's beliefs about teacher quality. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table A.3: Relationship between Teacher Value-Added and Characteristics

	Teacher Baseline Value-Added (in Student SDs)		
	(1)	(2)	(3)
Risk lovingness (coin flip game)	0.0139		
	(0.00988)		
Pro-sociality (volunteer task)		-0.00479	
		(0.00650)	
Dislike competition			-0.000677
			(0.00632)
Observations	5585	5585	5585
Control Mean	-0.0283	-0.0283	-0.0283
Control SD	0.349	0.349	0.349
Observations	5585	5585	5585

Notes: This table presents the relationship between teacher characteristics and baseline value-added controlling. *Teacher Baseline Value-Added* is measure of teacher value-added using test score data from the two years prior to the intervention. It is in student standard deviations. Characteristics (*risk lovingness*, *pro-sociality* and *dislike competition*) are measured in z-scores and collected at baseline. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table A.4: Sorting Controlling for Teacher Characteristics

	Teacher Baseline Value-Added (in Student SDs)			
	(1)	(2)	(3)	(4)
Chose Performance Pay	0.0485**	0.0467**	0.0494**	0.0486**
	(0.0207)	(0.0207)	(0.0207)	(0.0207)
Risk lovingness (coin flip game)		0.0126		
		(0.00990)		
Pro-sociality (volunteer task)			-0.00572	
			(0.00654)	
Dislike competition				-0.00190
				(0.00643)
Control Mean	-0.0283	-0.0283	-0.0283	-0.0283
Control SD	0.349	0.349	0.349	0.349
Observations	1284	1284	1284	1284

Notes: This table presents the relationship between teacher contract choice and baseline value-added controlling for teacher characteristics. *Teacher Baseline Value-Added* is measure of teacher value-added using test score data from the two years prior to the intervention. It is in student standard deviations. *Chose Performance Pay* is a dummy variable for whether a teacher chose performance pay or flat pay during the baseline choice exercise. Characteristics (*risk lovingness*, *pro-sociality* and *dislike competition*) are measured in z-scores and collected at baseline. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table A.5: Baseline Covariates - Neighboring School's Treatment

Variable	(1) Same Treatment as Teacher's School N/[Clusters]	Mean/SE	(2) Opposite Treatment as Teacher's School N/[Clusters]	Mean/SE	T-test Difference (1)-(2)
Performance evaluation score	2201	3.381	769	3.347	0.034
	[121]	(0.015)	[80]	(0.032)	
Salary (USD)	3026	5423.244	1018	5325.916	97.328
	[126]	(103.000)	[83]	(155.855)	
Age	3027	36.641	1018	37.096	-0.455
	[126]	(0.359)	[83]	(0.410)	
Years of experience	3020	5.756	1017	5.722	0.035
	[126]	(0.199)	[83]	(0.247)	

Notes: This table summarizes teacher and student characteristics before the experiment by neighboring schools treatment. The table reports mean values of each variable for each treatment group. The final three columns report mean differences between treatment group. Panel A presents teacher demographics as of September 2017. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table A.6: Treatment Effect by Contract Choice, Across Question Type

	Endline Test (z-score)			
	All questions	External	Remedial	Advanced
Perf Pay Treat	0.00857	0.0424	0.0684	0.103
	(0.0511)	(0.0651)	(0.0910)	(0.112)
Chose Perf Pay	-0.0397	-0.0425	-0.0799	-0.000425
	(0.0338)	(0.0345)	(0.0529)	(0.0835)
Chose Perf Pay*Perf Pay Treat	0.0822**	0.0659	0.0939	0.0932
	(0.0406)	(0.0416)	(0.0692)	(0.114)
β(Treat + Treat*ChosePP)	0.09	0.11	0.16	0.20
pval(Treat + Treat*ChosePP)	0.03	0.04	0.01	0.03
Control Mean	7.94e-10	-0.0314	-0.0499	-0.0667
Control SD	1.000	1.007	1.015	1.023
Clusters	114	113	100	90
Observations	144009	102739	40560	19487
Randomization Strata FE	Yes	Yes	Yes	Yes
Baseline	Yes	Yes	Yes	Yes

Notes: This table presents the treatment effect of performance pay contracts on endline tests scores by contract choice. *Perf Pay Treat* is a dummy for whether a teacher taught at a school assigned to performance pay versus flat pay school at baseline. *Chose Perf Pay* is a dummy variable for whether a teacher chose performance pay or flat pay during the baseline choice exercise. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table A.7: Treatment Effects on Classroom Observations by Contract Choice

	CLASS Rubric				Test Prep
	All	Class Climate	Differentiation	Student-Centered	Minutes
Obj PP Treat	-0.409**	-0.473***	0.0131	-0.469***	0.283***
	(0.157)	(0.165)	(0.0919)	(0.165)	(0.0927)
Chose Obj PP	-0.124*	-0.0864	-0.112	-0.108	0.101
	(0.0727)	(0.0556)	(0.0754)	(0.0731)	(0.104)
Obj PP Treat*Chose Obj PP	0.568***	0.565***	0.338***	0.530***	-0.0737
	(0.131)	(0.130)	(0.0853)	(0.135)	(0.120)
β(Treat + Treat*ChosePP)	0.16	0.09	0.35	0.06	0.21
pval(Treat + Treat*ChosePP)	0.09	0.35	0.00	0.51	0.01
Control Group Mean	-0.05	0.03	-0.21	0.00	-0.17
Clusters	71	71	71	71	71
Observations	1956	1956	1956	1956	1956
Randomization Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes
Observer FE	Yes	Yes	Yes	Yes	Yes

Notes: This table presents the treatment effect of performance pay contracts on classroom observation scores by contract choice. *Obj PP Treat* is a dummy for whether a teacher taught at a school assigned to an objective performance pay versus flat pay school at baseline. *Chose Obj PP* is a dummy variable for whether a teacher chose objective performance pay or flat pay during the baseline choice exercise. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

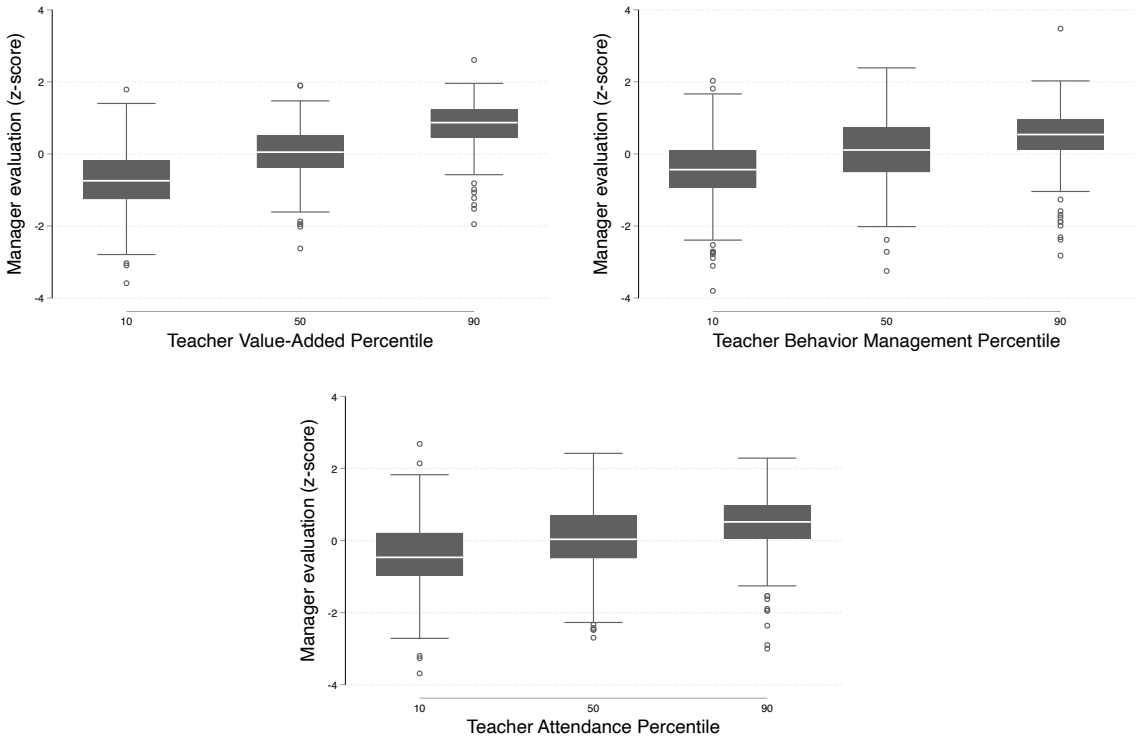
Table A.8: Treatment Effects on Student Survey by Contract Choice

	Endline Survey Indices (z-score)					
	All	Love of learning	Ethical	Global	Inquisitive	Dislike school
Obj PP Treat	0.0523	-0.0394	0.133	0.186	-0.144**	-0.0664
	(0.0380)	(0.0710)	(0.109)	(0.133)	(0.0658)	(0.0662)
Chose Obj PP	-0.0323	-0.0155	0.00178	-0.0661*	-0.0400	0.0171
	(0.0206)	(0.0263)	(0.0273)	(0.0354)	(0.0425)	(0.0172)
Obj PP Treat*Chose Obj PP	0.0645***	0.0506	0.0795	-0.0623	0.118*	-0.0462
	(0.0230)	(0.0596)	(0.0955)	(0.0871)	(0.0604)	(0.0344)
β(Treat + Treat*ChosePP)	0.12	0.01	0.21	0.12	-0.03	-0.11
pval(Treat + Treat*ChosePP)	0.00	0.86	0.01	0.10	0.77	0.03
Control Group Mean	-0.04	-0.09	-0.14	-0.02	-0.02	0.34
Clusters	33	33	33	33	33	31
Observations	16059	16046	16059	16029	16059	14291
Randomization Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table presents the treatment effect of performance pay contracts on student survey scores by contract choice. *Obj PP Treat* is a dummy for whether a teacher taught at a school assigned to an objective performance pay versus flat pay school at baseline. *Chose Obj PP* is a dummy variable for whether a teacher chose objective performance pay or flat pay during the baseline choice exercise. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

A.2 Supplementary Chap 2 Figures and Tables

Figure A.11: Manager Rating by Vignette Characteristics



Notes: These figures present box plots of principal's responses to vignettes asking managers to rate a hypothetical teacher based on a description of their performance. The vignettes stated, "[Female name/Male name] is in the [bottom/middle/top] 10% of teachers in terms of students' test score growth, in the [bottom/middle/top] 10% of teachers in terms of behavioral management, and is in the [bottom/middle/top]10% in terms of attendance and timeliness at work." Managers rated three such vignettes with characteristics randomized across vignettes. *Teacher Value-Added Percentile*, *Teacher Behavioral Management Percentile*, and *Teacher Behavioral Management Percentile* takes values, 10, 50 and 90 to correspond to the bottom, middle and top 10% listed in the vignette. *Manager evaluation (z-score)* is the residualized value of the manager's survey response, controlling for the three other characteristics. For example, in the first figure plotting *Teacher Value-Added Percentile* versus *Manager evaluation (z-score)* . Manager rating is residualized by *Teacher Behavioral Management Percentile*, *Teacher Behavioral Management Percentile* and *Female name*.

Table A.9: Percent of Time Individuals Believe Should be Spent on Each Type of Activity

Variable	(1) Objective Teachers N/ [Clusters]	(1) Objective Teachers Mean/ SE	(2) Subjective Teachers N/ [Clusters]	(2) Subjective Teachers Mean/ SE	(3) Subjective Managers N/ [Clusters]	(3) Subjective Managers Mean/ SE	T-test Difference (1)-(2)	T-test Difference (1)-(3)	T-test Difference (2)-(3)
Improving behavioral management	487	0.062	2406	0.059	41	0.054	0.003**	0.009*	0.006
		(0.001)		(0.001)		(0.006)			
Collaborating with other teachers	487	0.051	2406	0.050	41	0.059	0.001	-0.008*	-0.009**
		(0.001)		(0.000)		(0.005)			
Grading student papers	487	0.068	2406	0.071	41	0.069	-0.003	-0.002	0.001
		(0.002)		(0.001)		(0.005)			
Providing differentiated lessons	487	0.068	2406	0.070	41	0.067	-0.003	0.000	0.003
		(0.002)		(0.001)		(0.005)			
Helping with extracurriculars	487	0.055	2406	0.056	41	0.047	-0.001	0.008	0.009
		(0.002)		(0.001)		(0.005)			
Incorporating higher order thinking skills	487	0.067	2406	0.067	41	0.067	0.001	-0.000	-0.001
		(0.002)		(0.001)		(0.005)			
Catering to different learning styles	487	0.066	2406	0.066	41	0.065	0.000	0.001	0.001
		(0.001)		(0.001)		(0.005)			
Incorporating multimedia	487	0.053	2406	0.056	41	0.053	-0.004**	-0.000	0.003
		(0.001)		(0.001)		(0.006)			
Communicating with parents	487	0.042	2406	0.040	41	0.042	0.002	0.001	-0.002
		(0.001)		(0.001)		(0.004)			
Conducting practice tests	487	0.067	2406	0.065	41	0.068	0.002	-0.001	-0.003
		(0.002)		(0.001)		(0.007)			
Making lessons more student centered	487	0.066	2406	0.070	41	0.083	-0.003**	-0.017***	-0.013***
		(0.001)		(0.001)		(0.007)			

Notes: This table reports teachers' responses to a hypothetical scenario in which they are advising a teacher which actions they should take to increase their raise under a given treatment. Data was collected as part of the endline survey, and observations are at the unit of the teacher/manager. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table A.10: Manager Rating by Vignette Teacher Characteristic

	Manager Rating (z-score)					
	(1)	(2)	(3)	(4)	(5)	(6)
Teacher Value-Added Percentile	0.0180***				0.0177***	0.0174***
	(0.00103)				(0.000979)	(0.00103)
Teacher Behavioral Management Percentile		0.00899***			0.00904***	0.00817***
		(0.000941)			(0.000724)	(0.000819)
Teacher Attendance Percentile			0.00791***		0.00805***	0.00738***
			(0.00121)		(0.000815)	(0.000917)
Teacher has female name				-0.0253	0.0166	0.0163
				(0.0557)	(0.0420)	(0.0477)
Constant	-0.885***	-0.451***	-0.389***	0.0128	-1.731***	-1.639***
	(0.0738)	(0.0684)	(0.0731)	(0.0634)	(0.106)	(0.0825)
Observations	567	567	567	567	567	567
Manager Fixed Effects						X

Notes: This table presents results from endline survey questions asking managers to rate a hypothetical teacher based on a description of their performance. The vignettes stated, "[Female name/Male name] is in the [bottom/middle/top] 10% of teachers in terms of students' test score growth, in the [bottom/middle/top] 10% of teachers in terms of behavioral management, and is in the [bottom/middle/top]10% in terms of attendance and timeliness at work." Managers rated three such vignettes with characteristics randomized across vignettes. *Teacher Value-Added Percentile*, *Teacher Behavioral Management Percentile*, and *Teacher Attendance Percentile* takes values, 10, 50 and 90 to correspond to the bottom, middle and top 10% listed in the vignette. *Teacher has female name* is a binary variable, which is 1 if the name used in the vignette is a traditionally female Pakistani name (Saadia, Haya, Maira, Anam, Zahra, or Sarah) and 0 if the name used is a traditionally male Pakistani name (Qasim, Tahir, Asim, Zain, Mujahid or Attefaq). Standard errors are clustered at the manager level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table A.11: Teacher Effort and Subjective Performance Rating

	Subjective Performance Rating Percentile		
	(1)	(2)	(3)
Hours present at school	-1.793***	-1.550***	-1.979***
	(0.293)	(0.306)	(0.617)
Days present at school	0.167***	0.153***	0.232***
	(0.0263)	(0.0263)	(0.0628)
Value-Added	1.393**	1.574***	3.417**
	(0.575)	(0.581)	(1.388)
CLASS Rubric Dimensions:			
Positive Climate			-7.472*
			(3.836)
Teacher Sensitivity			0.323
			(3.647)
Regard for Student Perspectives			1.650
			(1.847)
Behavioral Management			0.282
			(3.574)
Productivity			-3.829
			(3.492)
Negative Climate			-12.49*
			(6.865)
Instructional Learning Formats			5.060
			(3.396)
Content Understanding			1.780
			(3.051)
Analysis and Inquiry			-4.815**
			(2.268)
Quality of Feedback			3.681
			(2.791)
Student Talk Time			-5.721**
			(2.427)
Student Engagement			5.804
			(3.651)
Other aspects of classroom observation:			
Students Use of English			-0.0498
			(0.0747)
Classroom is decorated			-0.659
			(6.348)
Use of technology			-0.0803
			(0.780)
Time spent on test prep			0.978
			(1.310)
Observations	2778	2628	618
Dependent Variable Mean	49.05	49.05	49.05
Subject and Grade Controls		X	X

Notes: This table presents the relationship between teacher behavior and their subjective performance rating. The dependent variable is subjective performance rating percentile. Column (1) and (2) includes the full sample of teachers and column (3) just includes teachers for whom we conducted a classroom observation. *Hours* and *Days present* are from biometric clock in and out data provided by the school system. Value-added is calculated using administrative test scores and endline test scores. The remaining variables are from classroom observations. The first 12 are the dimensions of the CLASS rubric and the rest are additional elements of teaching not captured by the CLASS rubric. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table A.12: Teacher's beliefs about contract features

Variable	(1) Objective Treatment N/[Clusters]	Mean/SE	(2) Subjective Treatment N/[Clusters]	Mean/SE	T-test Difference (1)-(2)
Panel A: Bias and Favoritism					
Is their any bias in favor or against the following groups (in the raises they receive)?					
New teachers	382 [33]	2.982 (0.029)	4237 [237]	2.983 (0.011)	-0.001
Female teachers	382 [33]	3.076 (0.029)	4237 [237]	3.077 (0.012)	-0.001
Older teachers	382 [33]	3.259 (0.054)	4237 [237]	3.248 (0.015)	0.011
Certain teachers are favored regardless of how hard they work	382 [33]	2.754 (0.050)	4237 [237]	2.772 (0.021)	-0.018
Panel B: Gaming					
Teachers do favors for managers to get a higher raise	124 [29]	2.427 (0.102)	2175 [208]	2.318 (0.038)	0.109
Teachers try to negotiate for a higher raise	124 [29]	2.548 (0.198)	2175 [208]	2.557 (0.037)	-0.009
Teachers bribe managers for a higher raise	124 [29]	1.508 (0.090)	2175 [208]	1.493 (0.026)	0.015
Panel C: Other features of the treatment					
How frequently did you think about the appraisal system	382 [33]	3.463 (0.149)	4237 [237]	3.479 (0.046)	-0.016
When did you come to understand what was expected under the contract	380 [33]	4.095 (0.128)	4199 [237]	4.089 (0.053)	0.006

Notes: This table summarizes teacher responses to questions about their contracts from the previous year at endline. The table reports mean values of each variable for objective versus subjective teachers. The final column report mean differences between treatment group and report if any are statistically significant. The three "Is there any bias" questions are on a 5 pt scale (1, lots of bias against, 3, no bias, 5, lots of bias in favor). The remaining questions in panel A and B are on a 5-pt scale from 1 (strongly disagree) to 5 (strongly agree). Questions in panel C were on a scale from 1 to 8. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table A.13: Heterogeneous Treatment Effects by Manager Characteristics

	Endline Test Scores					
	(1)	(2)	(3)	(4)	(5)	(6)
Subjective Treatment	-0.0156	0.169**	-0.0566	0.249***	0.412	-0.0912
	(0.197)	(0.0688)	(0.117)	(0.0775)	(0.681)	(0.491)
Interaction	0.00111	0.00827	0.0159	0.142*	0.0386	-0.0215
	(0.00274)	(0.00503)	(0.0977)	(0.0763)	(0.0863)	(0.0618)
Interaction*Subjective Treatment	0.00205	-0.00883	0.148	-0.211**	-0.0818	0.0375
	(0.00420)	(0.00648)	(0.127)	(0.0910)	(0.162)	(0.116)
Interaction	Age	Experience (years)	Female	Manager innacuracy (z-score)	Management Rating	Personnel Management Rating
Clusters	255	255	255	255	255	255
Observations	440595	440595	440595	440595	440595	440595

Notes: This table presents the treatment effects by manager characteristics. The row *Interaction* lists which characteristic is used as the interaction variable for a given column. Age, experience and gender are from administrative records. Manager inaccuracy is from teacher endline survey data. Mangement rating and Personnel management rating are from manager endline survey responses to World Management Survey questions. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

A.3 Proofs

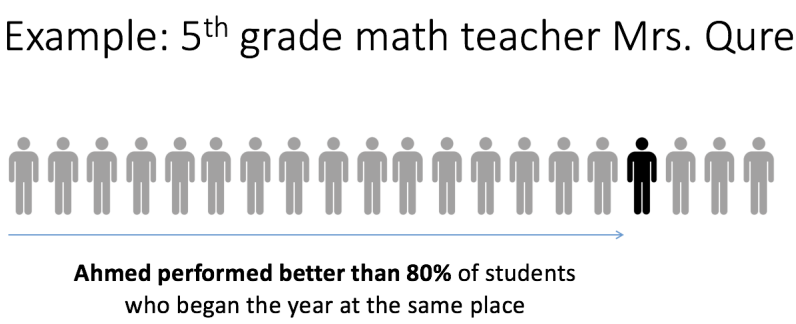
Proof of eq. 1.5

This proof demonstrates how the total effect of offering a performance pay contract is the sorting effect on ability, sorting effect on treatment effect, and the average treatment effect. Here p is the fraction of individuals for whom $b \geq 0$.

$$
\begin{aligned}
\Delta y &= E[\theta + \beta | b \geq 0] - E[\theta | b < 0] \\
&= E[\theta | b \geq 0] - E[\theta | b < 0] + E[\beta | b \geq 0] && \text{linearity of expectation} \\
&= E[\theta | b \geq 0] - E[\theta | b < 0] + E[\beta | b \geq 0] + (-E[\beta] + E[\beta]) \\
&= E[\theta | b \geq 0] - E[\theta | b < 0] + E[\beta | b \geq 0] - (E[\beta | b \geq 0]p + E[\beta | b < 0](1-p)) + E[\beta] && \text{def. of expectation} \\
&= E[\theta | b \geq 0] - E[\theta | b < 0] + (E[\beta | b \geq 0] - E[\beta | b < 0])(1-p) + E[\beta] && \text{re-grouping} \\
&= E[\theta | b \geq 0] - E[\theta | b < 0] + (E[\beta | b \geq 0] - E[\beta | b < 0])P(b < 0) + E[\beta] && \text{law of total probability}
\end{aligned}
$$

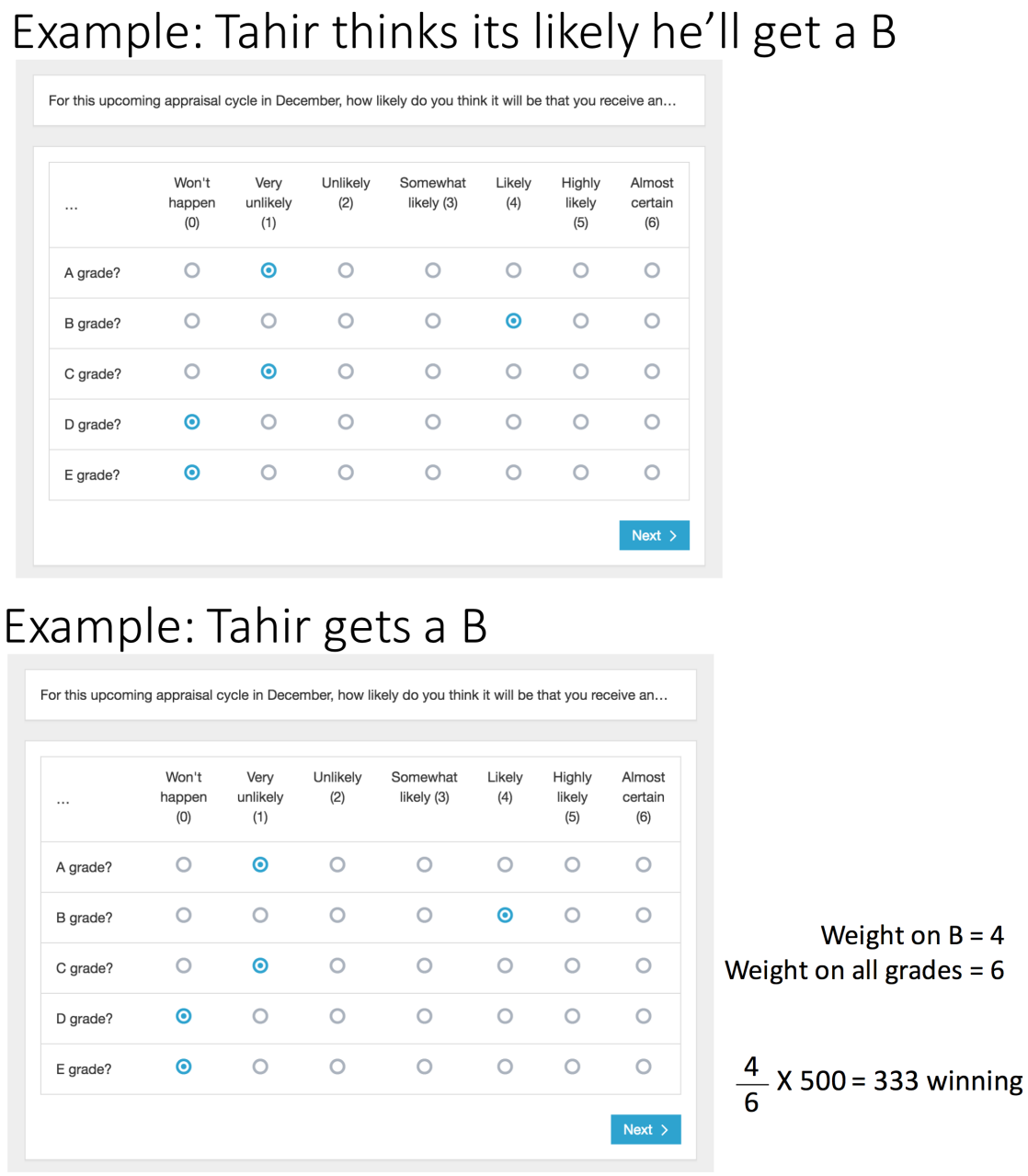
A.4 Experimental Design Implementation

Figure A.12: Screen capture from survey video: Calculation of percentile VA



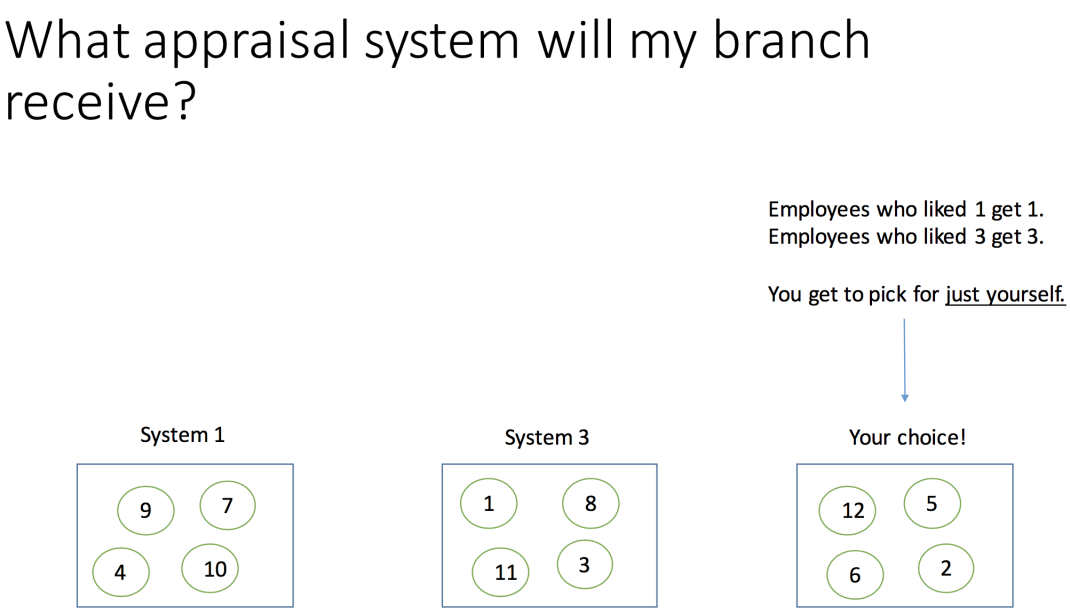
Notes: Screen capture from the video explaining to teachers how percentile value-added was calculated, giving teachers practical examples.

Figure A.13: Screen capture from baseline survey: Incentivized belief distribution elicitation



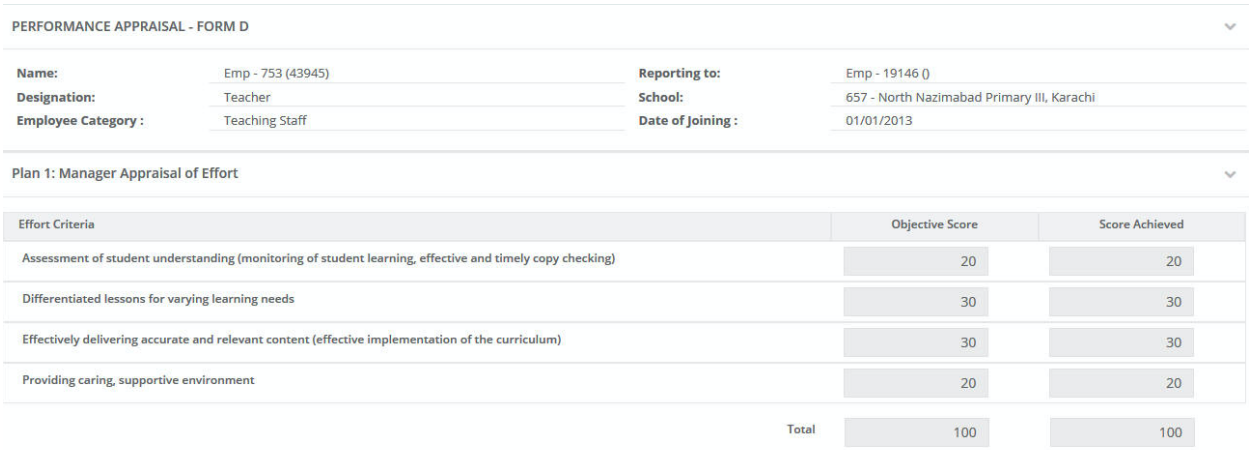
Notes: These figures are two screen shots from the video explaining to teachers how they would be incentivized for their beliefs about their value-added. Teachers are already familiar with this "A grade", "B grade" language which is used internally to rank teachers and captures teacher percentile. We borrow that same terminology for the survey questions since teachers are very familiar with it.

Figure A.14: Screen capture from baseline survey: Contract randomization



Notes: This figure shows a screen capture from the video explaining to teachers how their contract choice would be implemented with some probability.

Figure A.15: Example Performance Criteria



PERFORMANCE APPRAISAL - FORM D

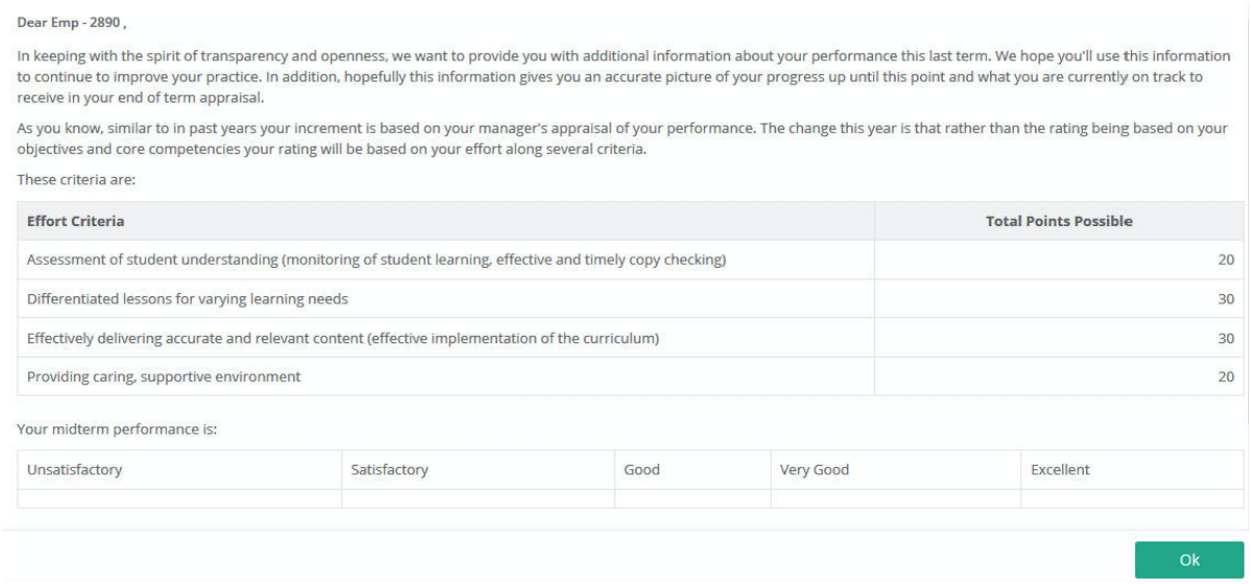
Name:	Emp - 753 (43945)	**Reporting to:**	Emp - 19146 ()
Designation:	Teacher	**School:**	657 - North Nazimabad Primary III, Karachi
Employee Category :	Teaching Staff	**Date of Joining :**	01/01/2013

Plan 1: Manager Appraisal of Effort

Effort Criteria	Objective Score	Score Achieved
Assessment of student understanding (monitoring of student learning, effective and timely copy checking)	20	20
Differentiated lessons for varying learning needs	30	30
Effectively delivering accurate and relevant content (effective implementation of the curriculum)	30	30
Providing caring, supportive environment	20	20
Total	100	100

Notes: This figure shows an example set of performance criteria a teacher would have set in collaboration with their manager at the beginning of the year. This list of criteria was located on their employment portal, and available to access throughout the year. Managers could set individual criteria for each of their employees. These ranged from 4 to 10 criteria spanning numerous aspects of the teacher's job descriptions.

Figure A.16: Example Midterm Information



Dear Emp - 2890 ,

In keeping with the spirit of transparency and openness, we want to provide you with additional information about your performance this last term. We hope you'll use this information to continue to improve your practice. In addition, hopefully this information gives you an accurate picture of your progress up until this point and what you are currently on track to receive in your end of term appraisal.

As you know, similar to in past years your increment is based on your manager's appraisal of your performance. The change this year is that rather than the rating being based on your objectives and core competencies your rating will be based on your effort along several criteria.

These criteria are:

Effort Criteria	Total Points Possible
Assessment of student understanding (monitoring of student learning, effective and timely copy checking)	20
Differentiated lessons for varying learning needs	30
Effectively delivering accurate and relevant content (effective implementation of the curriculum)	30
Providing caring, supportive environment	20

Your midterm performance is:

Unsatisfactory	Satisfactory	Good	Very Good	Excellent

Ok

Notes: This figure shows an example notification sent to teachers during the summer between the two school years. The notification gave teachers a preliminary performance rating based on the first term of the experiment. Teachers received this information via email and as a pop-up notification on their employment portal. This example shows the notification that subjective treatment teachers would receive. Teachers in the objective treatment received midterm performance information based on their students percentile value-added from the first term. Teachers in the control schools received information about either their performance along the subjective criteria that by their manager or their students' percentile value-added.

Table A.14: Socio-Emotional Outcomes Student Survey

Question	Category	Source
1. I enjoy my math/science/English/Urdu class	Love of learning	National Student Survey
2. When work is difficult, I either give up or study only the easy part (reversed)	Love of learning	Learning and Study Strategies Inventory
3. I get very easily distracted when I am studying or in class (reversed)	Love of learning	Learning and Study Strategies Inventory
4. I can spend hours on a single problem because I just can't rest without knowing the answer	Love of learning	Big Five (childrens)
5. I feel sorry for other kids who don't have toys and clothes	Ethical	Eisenberg's Child-Report Sympathy Scale
6. Seeing a child who is crying makes me feel like crying	Ethical	Bryant's Index of Empathy Measurement
7. It is ok if a student lies to get out a test they are worried about failing (reversed)	Ethical	
8. The pressure to do well is very high, so it is ok to cheat sometimes (reversed)	Ethical	
9. I am interested in public affairs	Global	Afrobarometer/World Values Survey
10. This world is run by a few people in power, and there is not much that someone like me can do about it (reversed)	Global	Afrobarometer
11. People who are poor should work harder and not be given charity (reversed)	Global	Afrobarometer
12. It is important to protect the environment even if this means we cannot consume as much today	Global	Afrobarometer
13. People from other places can't really be trusted (reversed)	Global	Afrobarometer
14. I am comfortable asking my math/science/Urdu/English teacher for help or support	Inquisitive	Learning and Study Strategies Inventory
15. I enjoy learning about subjects that are unfamiliar to me.	Inquisitive	Litman and Spielberger, Epistemic Curiosity questionnaire
16. I would like to change to a different school	Dislike school	Learning and Study Strategies Inventory

Notes: This table presents the student survey question items used to assess student socio-emotional skills. Students rated these questions on a 5-pt scale from Strongly disagree to Strongly agree.

Table A.15: Teacher Characteristics - Survey Items

Question	Category	Item Source
1. When it comes right down to it, a teacher really can't do much because most of a student's motivation and performance depends on students' home environment (reversed)	Efficacy	RAND Teacher Efficacy Index
2. If I really try hard, I can get through to even the most difficult or unmotivated students	Efficacy	RAND Teacher Efficacy Index
3. "Smartness" is not something you have, rather it is something you get through hard work	Efficacy	RAND Teacher Efficacy Index
4. A teacher is very limited in what he/she can achieve because a student's home environment is a large influence on the student's achievement (reversed)	Efficacy	RAND Teacher Efficacy Index
5. When a student gets a better grade than he usually gets, it is usually because I found better ways of teaching that student	Efficacy	RAND Teacher Efficacy Index
6. I expect to be in a higher-level job in five years	Career concerns	Ashraf et. al. (2020)
7. I view my job as a stepping stone to other jobs	Career concerns	Ashraf et. al. (2020)
8. I expect to be doing the same work as a teacher in five years (reversed)	Career concerns	Ashraf et. al. (2020)
9. Supporting students makes me very happy	Pro-social motivation	
10. I have a great feeling of happiness when I have acted unselfishly	Pro-social motivation	Ashraf et. al. (2020)
11. When I was able to help other people, I always felt good afterward	Pro-social motivation	Ashraf et. al. (2020)
12. Helping people who are not doing well does not raise my own mood (reversed)	Intrinsic Motivation (pro-social)	Ashraf et. al. (2020)
13. It is important to me to do good for others through my work	Intrinsic Motivation (pro-social)	Ashraf et. al. (2020)
14. I want to help others through my work	Intrinsic Motivation (pro-social)	Ashraf et. al. (2020)
15. One of my objectives at work is to make a positive difference in other people's lives	Intrinsic Motivation (pro-social)	Ashraf et. al. (2020)
16. The people, such as students or other teachers, who benefit from my work are very important to me	Intrinsic Motivation (pro-social)	Ashraf et. al. (2020)
17. My students matter a great deal to me	Intrinsic Motivation (pro-social)	Ashraf et. al. (2020)

Notes: This table presents the teacher survey question items used to assess teacher characteristics. Teachers rated these questions on a 5-pt scale from "Strongly disagree" to "Strongly agree". Items 9, 16 and 17 were adapted from their original language to refer to helping "students" rather than the generic "people", which is the phrasing in the original study.