Inducing Positive Sorting through Performance Pay: Experimental Evidence from Pakistani Schools[*]

Christina Brown[†] Tahir Andrabi

February 3, 2025

Abstract

Attracting and retaining high-quality teachers has large social benefits, but schools often struggle to attract and retain good teachers. This paper uses an experiment with 7,000 teachers in Pakistan to study whether performance pay affects not only teacher effort but selection. Consistent with adverse selection models, we find performance pay induces positive sorting both among teachers with higher initial ability and those with larger effort response to performance pay. Using two additional treatments, we show effects are larger among teachers with more private information about their ability and lower job switching costs. These findings demonstrate that the total effect of performance pay on test scores is substantially larger when including these employee composition effects.

[*] We are incredibly thankful for support and advice from Supreet Kaur, Christopher Walters, and Edward Miguel. David Card, Jishnu Das, Stefano DellaVigna, Frederico Finan, Anne Karing, Asim Khwaja, Samuel Leone, Patrick Kline, Jeremy Magruder, Peter McCrory, Preston Mui, Gautam Rao, Jesse Rothstein, Heather Schofield, Peter Serneels, and numerous seminar and conference audiences provided helpful feedback. Funding for this project was provide by the UK Foreign, Commonwealth & Development Office, awarded through the J-PAL Post-Primary Education Initiative and RISE Programme, the Weiss Family Fund, CEGA and the Strandberg Fund. Christina acknowledges support from the National Academy of Education/Spencer Dissertation Fellowship and the Institute for Research on Labor and Employment Fellowship. Our wonderful team at the Center for Economic Research in Pakistan, Haya Mubasher, Anam Tariq, Attefaq Ahmed, Zohaib Hassan, Zahra Niazi, Ayman Moazzam, Mujahid Murtaza and Maheen Rashid, provided excellent research assistance. All remaining errors are our own. We received IRB approval from Pomona College. AEA Registry-0004471.

[†] Brown (corresponding author): University of Chicago (christinabrown@uchicago.edu); Andrabi: Pomona College

Understanding how organizations can attract and retain high quality employees is a fundamental question for firms and bureaucracies. In a global survey of firms, managers state identifying, hiring and retaining high-quality employees is one of the most pressing problems they face (The World Bank , 2019). This need is especially dire in professions where job security is high or screening ability is low. Performance pay contracts offer a potential solution to this adverse selection problem. Even if employers cannot identify employee quality directly at the time of hiring, high performers should sort into firms that offer performance pay if two conditions are met: i). employees have private information about their ability and ii). switching costs are sufficiently low (Akerlof, 1970; Lazear and Moore, 1984).

We test whether performance pay can indeed help correct adverse selection in teaching. This is a setting where employers typically struggle to effectively screen candidates, as 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). Understanding whether performance pay may have positive (or negative) selection effects is crucial from a policy perspective as performance incentives have become increasingly common in teaching. Currently, two-thirds of countries offer some form of performance incentives to public school teachers (The World Bank Group, 2018a). While there is a substantial body of evidence on how performance pay affects existing teachers' effort (Lavy, 2009; Muralidharan and Sundararaman, 2011; Fryer, 2013; Goodman and Turner, 2013), we know much less about whether performance pay changes *who* teaches.

In this paper, we use a large-scale experiment to answer the following questions: Can performance pay help schools attract and retain better teachers (in addition to increasing effort) and under what conditions? If so, how much private information do teachers have about their quality relative to what their manager knows? To answer these questions we partner with a network of over 200 private schools across Pakistan to vary the contracts offered and observe worker sorting and effort in response. Our experiment is informed by a Roy-style model of job choice in which different employers offer different contracts, and employees choose where to work based on their beliefs about their own ability and idiosyncratic preferences.

Our experiment proceeds in two phases. We begin at baseline by asking teachers to choose which contract they would prefer for the coming year, selecting between a flat raise versus a performance-based raise contract. To ensure incentive compatibility of responses, teachers are told that in a random subset of schools, everyone will receive the contract they chose in the survey. We also measure several aspects of teacher preferences, beliefs and characteristics, including risk, pro-sociality, career ambition, and beliefs about their ability rank at baseline.

Next, among the remaining schools that were not randomized to implement the teacher's contract choice, we randomize schools to one of three contracts. Under the flat raise contract (control), teachers are guaranteed a fixed raise of 5% irrespective of their performance. Under the performance raise contract (treatment), a teacher's raise varies from 0-10% based on their students' test scores (in a random subset of schools) or their principal's evaluation (in the other subset of schools). Then over the subsequent year, we observe what types of teachers move into schools assigned flat versus performance raise contracts. This two step design of first eliciting preferences and then randomizing contracts across schools allows us to compare the effect of being assigned a performance raise contract separately for teachers who wanted this contract versus did not.

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 effects on student test score and non-test score outcomes and teacher behavior, including daily clock-in and out time and rich measures of classroom practices using the *CLASS* classroom observation rubric.

Our conceptual framework yields four predictions which we are able to test with our experiment. First, if employees have non-zero private information about their "ability" (employee's output in the absence of incentives), firms offering performance pay will attract and retain higher ability individuals. Second, if employees have non-zero private information about their "behavioral response" (the increase in their output under performance pay versus flat pay), firms offering performance pay will also attract and retain employees with a larger behavioral response. Finally, the extent of this positive sorting on ability and behavioral response depend on the extent of private information employees have and how costly it is to switch into their preferred pay contract.

First, we find strong evidence that performance pay induces positive sorting – both among teachers with higher baseline *ability* and among teachers with a larger *behavioral response* to performance pay. We show evidence of positive sorting on ability in two ways: i). teachers who chose performance versus flat raise contracts have much higher baseline value-added and ii).teachers who work at schools randomly assigned to performance raise versus flat raise contracts a year after randomization also have higher baseline value-added. These effects are mostly driven by existing high value-added teachers moving from control to treatment schools and vice versa, rather than differences in who is entering this system of schools.

Second, to test whether there is positive sorting on teacher's behavioral response to incentives, we compare the effect of being assigned to a performance pay contract versus a flat pay contract for teachers who stated they preferred the performance pay contract in the baseline survey. The effect of being assigned to a performance pay contract relative to flat pay on test scores is nine times larger for those who wanted a performance pay contract than for those that wanted flat pay.

Moreover, the behavioral response is not correlated with baseline value-added, suggesting that these two aspects of teacher type, ability and behavioral response, are orthogonal. 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. This suggests that previous estimates from the literature which did not allow for compositional effects may substantially underestimate the effects of performance pay.

However, these findings alone do not tell us whether employees know more about their own ability and behavioral response than their manager. To do this, we add in our detailed measures of principal beliefs and other aspects observable to school administrators. We find the coefficient does not change when adding these controls, suggesting teachers have substantially more information about their ability than their managers. Teacher's contract decisions are three times as predictive of value-added as all the information available to schools (credentials, experience, age, and principal evaluation).

Principals do have some information about teacher quality, though, and they are especially good at rating teachers along highly observable criteria, such as attendance and behavioral management of students. The private information gap between what teacher's know about their own ability and what manager's know about the teacher's ability is largest in the middle of their careers (6-10 years of experience) suggesting it takes time for teachers to learn about their own ability.

Next, in order to understand how much sorting we would expect to see under alternative policies (e.g. all public schools offering this contract, schools introducing this contract for a longer time period, etc), we need to test how sensitive these effects are to the extent of private information teachers have and the magnitude of job switching costs. We use two additional sources of random variation to show that the extent of positive sorting varies substantially by the accuracy of teacher's priors and how costly it is for them to switch jobs. To vary the accuracy of teachers' priors, we provide information to a random subset of teachers about their value-added in the previous year. This information treatment doubles the extent of positive sorting. To vary job switching costs, we compare teachers whose closest neighboring school received the opposite treatment status as their own school (low switching cost) versus the same treatment status (high switching cost). The extent of positive sorting is four times higher for teachers with low switching costs.

Finally, we show that performance pay does not generate sorting of "bad" types into performance pay schools. Teachers who chose performance pay at baseline are much *less* likely to exhibit distortionary behaviors, like teaching to the test, in response to performance incentives than those who chose flat pay. They are also slightly more likely to contribute to school public goods, to collaborate with other teachers, and have similar levels of pro-sociality (measured using a volunteering task). This suggests that teachers who sort in are not solely focused on maximizing their salary at the cost of more well-rounded student development.

In the concluding section, we take a step back to understand what can be learned from this controlled experiment to understand the extent of sorting there may be if an entire school district or the entire teaching profession moved to performance pay. We use the estimates of teacher's private information, distribution of ability and behavioral response, and elasticity of supply to a given job from our experiment to estimate the effects of a long-term performance pay policy applied to all schools. We find that introducing a 30-year performance pay policy in an entire school district (occupation) would result in effects on student learning between 0.10 - 0.21SD (0.09-0.17 SD) each year. These effects are much larger than if we had just considered the direct behavioral effect of performance pay on the existing pool of teachers.

Our paper makes three key contributions. We provide the first empirical evidence that performance pay contracts induce positive sorting among incumbent teachers (through sorting across schools and exit). The closest paper in this literature is Leaver et al. (2021) which studies the extent of teacher sorting in response to district-level variation in job postings (performance vs. flat pay) for new teachers and then re-randomizes performance or flat pay at the school level a la Karlan and Zinman (2009).

In contrast our paper uses the experiment to measure the underlying sorting primitives (teacher private information and job switching costs) and focuses on sorting across jobs rather than across occupations. While Leaver et al. (2021) does not find much evidence of positive sorting, we find sizeable sorting effects. However, this is unsurprising when we consider the differences in the two contexts: they primarily focus on sorting at the occupation level for new entrants which is a likely a "high cost" switch. Both sets of findings are entirely consistent with the model we present and mirror our heterogeneity findings – no sorting effects when switching costs are high (section 6.2). Our findings are also consistent with other work showing different contract features can help schools attract and retain better teachers (Dee and Wyckoff, 2015; Johnston, 2020; Biasi, 2021).

Second, we provide the first empirical evidence in any sector of performance pay inducing positive selection by the extent of an employee's behavioral response. While there is a robust literature on the direct, behavioral effect of performance pay for the existing stock of teachers, our paper is the first to show effects vary substantially by whether the teacher wanted that contract type (Lavy, 2009; Muralidharan and Sundararaman, 2011; Fryer, 2013; Goodman and Turner, 2013). The possibility of adverse selection operating along the margin of an individual's extent of moral hazard has been discussed in the context of health insurance

(Finkelstein and McGarry, 2006) but has been largely absent from the empirical performance pay literature.

These findings suggest in the long run, the purely behavioral effects of performance incentives could be much larger than the short term effects previously estimated, due to more responsive types sorting into the profession. In addition, the behavioral response appears to be uncorrelated with baseline value-added. This suggests that the marginal effort response to incentives is uncorrelated with the equilibrium effort under no incentives.

Finally, our design allows us to measure the underlying adverse selection primitives (Akerlof, 1970; Lazear and Moore, 1984; Greenwald, 1986; Rothstein, 2015), so we can estimate the extent of sorting under a variety of policy regimes. In order to do this you need to know: i). employee's private information about performance pay dimension (employee information – manager information), ii). job or occupational switching costs, iii). existence of other relevant personnel policies (hiring, firing, etc) and the extent of manager information at these points. We estimate these by eliciting contract preferences and manager information for all teachers. Just looking at ex-post sorting in response to a contract change would not be sufficient to estimate these primitives as the observed sorting decision will depend jointly on both private information and switching costs. Our paper is the first to estimate i.-iii. in response to a performance pay contract (not just in teaching, but in any profession). The goal of the paper is not to solely evaluate this specific program, but to also measure something more general about the extent of information employees and employers have, and as a result how employees may self-select under alternative policy regimes.

The remaining sections are organized as follows: Section 1 provides context about the teacher labor market and use of performance pay in teaching. Section 2 presents the motivating model in the vein of Roy (1951). Section 3 details the contract choice elicitation, randomized controlled trial, and data collection procedures. Section 4 presents the results on the extent of positive sorting in response to performance pay, and section 5 describes the extent of information principals have about teachers. Section 6 presents results on the sensitivity of the magnitude of positive sorting to teacher's switching costs and information, and section 7 examines whether there is sorting along negative characteristics. Section 8 discusses how the framework and results for this study can help us make sense of findings from several related papers in this field and simulates counterfactuals for policies in which an entire district or the entire teaching profession moved to performance pay.

1 Teacher 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 (The World Bank Group, 2018b). The dearth of good teaching has large, long-lasting, and diverse negative consequences for students (Chetty et al., 2014; Jackson, 2018; Rose et al., 2022). In Pakistan, exposure to a one standard deviation (SD) better teacher results in 0.15 SD higher test scores (Bau and Das, 2020).

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. 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). However, 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 manufacturing employees positively sort in response to performance pay, and the total effects are twice as large when you account for sorting, rather than just effort response.

It is unclear whether we would see more or less asymmetric information in teaching, relative to manufacturing. On the one hand, it is likely harder for employers to assess productivity in higher-skilled professions, like teaching, which has a complicated production function. On the other hand, teacher performance pay is generally constructed using an opaque performance incentive metric (typically value-added), and teachers may struggle to assess their ability along this metric. In fact, 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 ask teachers to predict their rank along the performance metric. We also find no relationship between teachers' predictions and actual performance (appendix figure A1). 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 behavioral effects on existing teachers and sorting effects on teacher composition 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 for which teachers' wages are affected by student outcomes or teacher effort doubled in the last decade, from one-third to two-thirds of countries (The World Bank Group, 2018a).

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.

2 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 show what average output for the firm would be if they offer performance pay versus flat pay.

2.1 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_θ^2 and σ_β^2, respectively, and covariance $\rho_{\theta,\beta}$. The wage from each contract is then:

$$w(\theta_i, \beta_i, j) = \begin{cases} w_0 & \textit{if } j = j_F \\ py_i = p(\theta_i + \beta_i) & \textit{if } j = j_P \end{cases} \tag{1}$$

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. For example, an α_θ of 1 is perfect information and an α_θ of 0 implies no private information about ability.

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 7.2, 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:

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

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

$$b_i = p(\hat{\theta}_i + \hat{\beta}_i) - \frac{p\hat{\beta}_i}{2} + \epsilon_{iP} - (w_0 + \epsilon_{iF}) \tag{3}$$

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

2.2 Employee Quality by Contract Offered

We treat employment as a one-sided job choice by the employee. Firms are unable to screen on θ or β.[2] However, employers can choose what contract they offer–a flat pay contract or performance pay contract. The average output per worker, $\bar{y}(j)$, by contract offered is:

$$\bar{y}(j) = \begin{cases} E[\theta_i | b_i < 0] & \textit{if } j = j_F \\ E[\theta_i + \beta_i | b_i \geq 0] & \textit{if } j = j_P \end{cases} \tag{4}$$

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 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:

$$\begin{aligned} \Delta\bar{y} &= E[\theta_i + \beta_i | b_i \geq 0] - E[\theta_i | b_i < 0] \\ &= \underbrace{E[\theta_i | b_i \geq 0] - E[\theta_i | b_i < 0]}_{\text{sorting on ability}} + \underbrace{(E[\beta_i | b_i \geq 0] - E[\beta_i | b_i < 0])P(b_i < 0)}_{\text{sorting on behavioral effect}} + \underbrace{E[\beta_i]}_{\text{avg. behavioral effect}} \end{aligned} \tag{5}$$

[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 exerted under performance pay is $c_i e^2 = c_i(\frac{p}{2c})^2 = \frac{p\hat{\beta}_i}{2}$.

[2] Section 1 shows this is an appropriate assumption in the teaching profession. However, we will also relax this constraint by presenting results controlling for principal information to mimic settings where principals can screen employees.

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. We will estimate all three terms individually through our experiment.

2.3 Model Predictions

The key predictions of the model are:

- *Prediction 1).* $\alpha_\theta > 0 \implies E[\theta_i|b_i \geq 0] - E[\theta_i|b_i < 0] > 0$
  If employees have any private information about ability, then there will be positive sorting on ability.

- *Prediction 2).* $\alpha_\beta > 0 \implies E[\beta_i|b_i \geq 0] - E[\beta_i|b_i < 0])P(b_i < 0) > 0$
  If employees have any private information about their behavioral response, then there will be positive sorting on behavioral response.

- *Prediction 3).* $\frac{\partial \Delta y_s}{\partial \alpha} > 0$
  Higher accuracy about type increases positive sorting.

- *Prediction 4).* $\frac{\partial \Delta y_s}{\partial \sigma^2_\epsilon} < 0$
  Higher variance in non-wage utility decreases positive sorting.

Appendix figure A2 demonstrates these comparative statics using simulated data. 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. In our experiment, that assumption is satisfied by randomizing performance versus flat pay contracts across schools, helping us to test predictions 1 and 2. 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 3 and 4, respectively.

3 Experimental Design

3.1 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 June 2017 to May 2019 with a private school chain that operates nearly 300 schools located across Pakistan. Figure 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 (described in detail in appendix D). 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. 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 took several steps before and during the survey 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 for a subset of teachers their decision would actually be implemented. Teachers had to pass understanding checks before they were allowed to make the contract choice. Finally, in the years prior to the study, teachers were subject to variable raises based on factors like participation in trainings, assisting in after-school activities and other aspects of performance (none of which were related to student learning), so they were familiar with some of the aspects of a performance-based contract.

Phase 2 – Contract Randomization: To measure the behavioral effects of performance pay, we randomize contracts at the school level to the remaining 243 schools that were not selected to implement the teacher's contract choice. The contracts applied to all core teachers (those teaching Math, Science, English, Urdu, and Social Studies) in grades 4-13 and all three have the same expected value. The contracts were:

1). **Control: Flat Pay** - Teachers receive a flat raise of 5% of their base salary.

2). **Treatment: Performance Pay** - Teachers receive a raise from 0-10% based on their within-school performance ranking. The expected value of this contract is also 5%.

Table 1 shows the distribution of raise values by within-school percentile.[3] There are two treatment sub-arms, which vary the performance metric used to evaluate teachers. Teachers are ranked within their school on either:[4]

2a. *Objective Performance:* Percentile value-added (Barlevy and Neal, 2012) averaged across all students they taught during the spring and fall term.[5]

2b. *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.[6]

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. We will present pooled results for the subjective and objective treatments together for most results as the sorting effects are very similar and show results by sub-treatment in the appendix.[7] Along all of our main sorting outcomes, we cannot reject equality of effects between the two sub-treatments.

3.2 Data

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

[3] Appendix C describes the implications of a tournament style performance pay system on sorting, comparing results to that of a linear contract such as the one used in section 2.1. We show, given moderate switching costs, the tournament-based contract can still induce sorting.

[4] 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 distribution of the incentive pay rates is equivalent across the two treatments (as shown in table 1). 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.

[5] 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.

[6] These items varied substantially across schools and teachers. Examples include: improved behavioral management of students, end of year exam score targets, helping plan an after-school event, and improving students' spoken English proficiency.

[7] Comparing teacher's effort response to subjective versus objective performance pay is subject of a companion paper (Andrabi and Brown, 2024)

Administrative data: The administrative data details employee job description, salary, performance review score, attendance, and demographics for June 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).

Endline student testing and survey: A grade-specific endline test was conducted in January to measure performance in Reading (English and Urdu), Math, Science, and Economics in grades 4-13 (distributions are shown in appendix figure A3).[8] The items were written in partnership with the school system's curriculum and testing department to ensure the appropriateness of question items. Items from international standardized tests (TIMSS and PERL) and a locally used standardized test (LEAPS) were also included to benchmark student performance. The research team conducted the grading. Students also completed a survey to measure four areas of socio-emotional development chosen based on the school system's student development priorities.[9] Appendix table D.2 lists the survey items used for each area along with their source.

Teacher and principal survey: In addition to the contract choice exercise, the baseline survey included incentivized measures of teachers' beliefs about their performance along the objective (percentile value-added) and subjective (principal evaluation) metric (appendix figure D.1c). We 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 school 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.[10] The survey also included measures of intrinsic motivation (Ashraf et al., 2020), efficacy (Burrell, 1994), and checks on what

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

[9] The areas are (i). love of learning (items drawn from National Student Survey, Learning and Study Strategies Inventory), (ii). ethics (items from Eisenberg's Child-Report Sympathy Scale, Bryant's Index of Empathy Measurement), (iii.) global citizenship (items from Afrobarometer; World Values Survey), and (iv.) inquisitiveness (items from Learning and Study Strategies Inventory; Epistemic Curiosity Questionnaire). These are the four areas of socio-emotional development teachers are expected to focus on. These areas are posted on the walls in schools, and teachers receive professional development in these topics. Some principals also specifically make these areas part of teachers' evaluation criteria.

[10] Teachers are asked about what size bonus they would need in order to transfer to a given school with a given performance or flat pay contract. To attach stakes to the question, we let teachers know that those with the lowest willingness to accept will be contacted first in case of opening and paid their stated amount (first price auction).

teachers understood about their assigned contract. The endline survey was conducted online with teachers and managers in spring and summer 2019. Appendix table D.1 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.

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. We also recorded whether teachers conducted any sort of test preparation activity and the language fluency of teachers and students.

3.3 Measuring Teacher Ability

To measure teacher's "ability", θ, we calculate teacher value-added (VA) using student test scores from June 2015, 2016 and 2017, the three years prior to the randomized controlled trial. This allows us to measure teacher effectiveness in the absence of performance pay. 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:

$$\begin{aligned} y_{ijkcst} = \beta_0 \quad & + \sum_s \beta_s y_{ijkcs,t\text{-}1} \mathbb{1}[\textit{subject-grade} = s] + \sum_s \alpha_s y_{ijkcs,t\text{-}2} \mathbb{1}[\textit{subject-grade} = s] \\ & + \sum_s \gamma_s \bar{y}_{\text{-}ijkcs,t\text{-}1} \mathbb{1}[\textit{subject-grade} = s] + \chi_{st} + \psi_k + v_{ijkcst} \qquad (6) \\ & \textit{where} \;\; v_{ijkcst} = \mu_j + c_{ct} + \epsilon_{ijkcst} \end{aligned}$$

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, c_{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).[11] Teachers for which we have few student observations are therefore shrunk

[11] 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

toward the mean teacher value-added (normalized to be zero).[12]

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 (2015) and 0.15 SD, Bau and Das (2020)). Appendix figure A4 shows the distribution of teacher value-added for our teachers at baseline.

3.4 Sample and Intervention Fidelity

Teacher and Principal Sample: The study was conducted with a large, high fee private school system in Pakistan. Table A1, 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 of teachers is mostly female (81%), young (35 years on average), with an average of 10 years experience, though a quarter of teachers are in their first year teaching. Yearly turnover within a school is 29%. Teachers are generally younger and less experienced than their counterparts in public schools, though they have more education and higher salaries. 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 skills than the average school in India. Given these differences these schools may be more equipped to implement performance pay effectively.

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 performance pay treatment arms and control appear balanced along baseline covariates. Appendix table A2 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. All

shrinkage factor.

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

results control for these unbalanced variables. Table A3 shows baseline balance in covariates for teachers along two other treatment dimensions used to test mechanisms in section 6.1 and section 6.2. Of eight tests, none are statistically significant.

Administrative data is available for all teachers and students who stay employed or enrolled during the year of the intervention. 88% of employed teachers completed the endline survey, and there was no differential attrition by treatment status. 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 differentially likely to opt-out across treatment. Finally, we find no evidence of differential student switching across schools in response to the treatments, which is unsurprising given the short intervention length. For longer term policies, we might expect student sorting if parents are able to observe differences in school quality as a result of contract type.

Teachers appear to understand their treatment assignment. On average, teachers stated that they came to fully understand what was expected of them by April 2018 (four months into the treatment year). However, knowledge of treatment status of other schools was relatively low, which may impede sorting across schools. Only 15% of teachers could name another school with a given treatment arm off the top of their head.

4 Positive Sorting

We now present the main results of the paper in sections 4 through 7. In this section, we present test predictions 1 and 2: whether performance pay induces sorting on ability and behavioral response. We test prediction 1 in two ways. We first test whether 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 test whether higher value-added teachers are more likely to move into performance pay schools after contracts have been randomized across schools. Finally, to test prediction 2 (whether there is sorting on behavioral response), we comparing the effect of being assigned a performance pay contract relative to flat pay for those that wanted the performance pay contract versus those that did not.

4.1 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?"*

We ask this question twice: once for an objective performance metric (percentile value-added) and once for a subjective performance metric (principal evaluation).[13] Appendix D provides screenshots from the video teachers watched about how percentile value-added is calculated and how teacher contract choices would be implemented.

Most teachers want at least part of their raise to be performance-based, with less than 10% choosing a completely flat raise. Figure A5 shows the distribution of teachers' responses. 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 simplicity, 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, we present results using the continuous measure in the appendix. All of the main results are unchanged between the two approaches.

Many aspects of teachers' beliefs, preferences and characteristics correlate with contract choice (figure A7). For example, teacher's belief about their principal's future rating of them is a strong predictor of contract choice. Teachers that are more risk-loving (as measured in a real-stakes coin flip game) and those that say they are likely to remain a teacher in the near future also prefer performance pay. Female teachers are less likely to choose performance pay, and experienced teachers are slightly more likely to choose performance pay. Teachers stated reason for choosing the given contract is shown in appendix figure A6.

Sorting on Ability – Contract Choice: We find that teachers who chose a performance pay contract have significantly higher baseline value-added. Figure 2 plots the distribution of baseline value-added (in student standard deviations) for teachers who chose performance pay

[13] 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.

(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 0.047 SD difference in student test scores. This difference holds for the choice between objective performance pay versus flat pay (0.049 SD difference) and subjective performance pay versus flat pay (0.045 SD difference). This relationship between teacher value-added and contract preference also remains when we consider a continuous measure of contract choice. Appendix figure A10 shows there is a strong relationship between value-added and the fraction of the teacher's raise they chose to be tied to performance.

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

$$VA_{i,t-1} = \beta_0 + \beta_1 ChosePP_i + \epsilon_i \tag{7}$$

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 2 presents the results from eq. 7. As we showed in the figures, teachers who chose performance pay had 0.05 SD higher baseline value-added. The relationship is similar whether we look at choices on objective or subjective performance pay. Columns (2), (4) and (6) 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, and the coefficient shrinks by just 9%. This suggests that teachers have substantial 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 A8 presents the relationship between baseline value-added and likelihood of choosing performance pay by teacher gender 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 and the relationship between ability and contract choice is slightly weaker than for male teachers. Mid-career teachers show the most sorting on ability, suggesting they have the best information about their own performance. On the other hand, novice teachers appear to have less information about their ability or, at least, are not sorting on that information. These results are consistent with Leaver et al. (2021) which does not find any sorting in response to performance pay among teachers entering the profession. Finally, very experienced teachers also do not exhibit sorting, and this appears to be driven by overconfident low-performers selecting the performance pay contract.

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 also observe if a teacher joins or leaves the school system but do not know if and where they are employed if they leave.[14,15] 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.[16] Transfers are initiated by the teacher and need to be accepted by the receiving school.[17] 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 was the set-up in the model, as the schools have little say in who within the transfer applicants is hired.

Sorting on Ability – Job Choice: Figure 3 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. The cumulative distribution lie on top of each other at baseline, but, a year later, the performance pay schools dominate flat pay

[14] 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.

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

[16] Teachers were provided information about other schools' treatment status over email and through their employee portal at the beginning of the intervention and again during the summer break, which is when most transfers take place.

[17] 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.

schools at every part of the distribution.

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

$$VA_{i,t-1} = \beta_0 \quad +\beta_1 WorkatPP_i + \beta_2 Post_i + \beta_3 WorkatPP_i * Post_i + \chi_j + \epsilon_i \tag{8}$$

$VA_{i,t-1}$ is the teacher's value-added at baseline. $WorkatPP$ 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 working in 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 3, column 1, presents the results of eq. 8. As we saw in 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 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 them). 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 from a more permanent contract change.

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 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.

Figure 4 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 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 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 flat pay schools had an average baseline value-added of 0.004 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 0.096 SD. Whereas, the average quality of those who moved from performance pay to flat pay is 0.004 SD. We also see better teachers leave the school system from flat pay schools (0.028 SD) than performance pay schools (-0.009 SD), which is consistent with positive sorting, but the difference is not statistically significant. About half of the teachers that leave the system go to work at another school and half get a non-teaching job or leave the labor market. We do not see significant differences in the quality of teachers who newly enter the system into a performance pay versus flat pay school (-0.031 SD vs -0.020 SD). This is not surprising, as the treatments were not advertised to new hires and were set to expire before new hires would begin receiving them (see Leaver et al. (2021) for a test of this type of sorting among first year teachers).

4.2 Positive Sorting on Behavioral Effect

Do teachers who chose performance pay also have a larger behavioral response? To test prediction 2, we compare the effect of being assigned a performance pay contract for those that wanted a performance pay contract versus flat pay in the baseline survey:

$$\begin{aligned} Y_i = \beta_0 \quad & +\beta_1 AssignedPP_j + \beta_2 ChosePP_i \\ & +\beta_3 AssignedPP_j \cdot ChosePP_i + \beta_4 Y_{i,t-1} + \chi_j + \epsilon_i \end{aligned} \tag{9}$$

The outcome, Y_i, is endline test scores for students taught by teacher, i. $AssignedPP_j$ 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 4.1, some teachers change schools during the experiment, so $AssignedPP_j$ 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, $Y_{i,t-1}$. 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, $AssignedPP_j$.

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 5 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 4, column 3, presents the results of equation 9. Column 4 controls for the principal's rating of the teacher, 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 A4, are the same if we treat contract choice as a continuous variable.

Is this "sorting on behavioral effect" just picking up the same high ability (baseline value-added) teachers who wanted performance pay? It does not appear that is the case. Column 5 shows there is no relationship between baseline value-added and behavioral effect. Column 6 shows that the coefficient on $AssignedPP_j \cdot ChosePP_i$ 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 necessarily the same individuals.

Total Effect of Performance Pay: Table 6 presents the total effect of performance pay adding together sorting on ability, sorting on behavioral effect and average behavioral effect. Column (1) and (2) show effect using the results from the contract choice exercise, which we can think of as zero switching costs. Column (3) and (4) shows the effect using the results from the teacher's job choice in the second year, which is a relatively high switching cost scenario.[18] The behavioral effect is 0.066 SD, coming from the average treatment effect of performance pay (table 4, col (2)). The fraction of the total effect which comes from sorting is 53% (low switching cost) and 33% (high switching cost). Our results are consistent with Lazear (2000) which found about half of the total effect of performance pay came from sorting and half from behavioral response in a manufacturing setting.

5 Asymmetric Information

5.1 How much information do employers have?

As we saw in table 2 and 3, 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, (measured via biometric clock in/out data), teachers' management of student discipline and incorporation of higher-order skills (measured using classroom observation data), and teachers' actual value-added.

Table 7 presents the relationship between principals' beliefs and teachers' actual outcomes. Pooling across all four dimensions (panel B, 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 (panel A, columns 1-4), 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 -0.05, respectively). More experienced principals are not any more accurate in rating teachers.

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 to conduct random spot checks of treated

[18] We cannot directly estimate the sorting on behavioral effect for the job choice experiment because we only observe teachers who move (or stay in) performance pay schools under that one contract. In order to calculate the total effect in column (3), we assume the ratio of sorting on ability to sorting on behavioral effect is the same under the contract choice and job choice and therefore extrapolate that the sorting on behavioral effect for the job choice experiment would be 0.011 SD.

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 this phase, relative to 1.8 for control.

Principals provide much more accurate ratings for teachers who were assigned to these unannounced classroom observations. Table 7, (panel B, column 2), 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 7, column 3, compares principal accuracy for principals who have worked at the same school as the teacher for more than or less than two years.[19] 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.[20] These effects are driven by principals boosting scores of low performing teachers the longer they overlap with them (figure A9).

Because overlap is not randomly assigned in this context, we cannot be sure if this relationship 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 4 controls for principal fixed effects and teacher's years of experience. Our results are robust to the addition of these controls. This provides suggestive evidence that simply sending time working with an employee is insufficient to close the private information gap.

5.2 How much more information do teachers have than managers?

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 A11 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

[19] 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.

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

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 A13 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 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.

What is the source of teacher's private information? There are two possible explanations for this result: (i) teachers have information about their own value-added 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. Our results are more consistent with the first explanation. Higher value-added teachers and those that have larger behavioral responses do not have different risk preferences, preferences for competition, or pro-sociality (figure A7). We can also control for risk preferences, preferences for competition, and pro-sociality in our main positive sorting results (table A5). Our results remain unchanged when we control for these potential channels.

6 Magnitude of Positive Sorting

Our experiment also allows us to explicitly test predictions 3 and 4, to see the effect of teacher's information and switching costs on the extent of positive sorting. First, we randomly provide some teachers with historical information about their value-added to test the effect of information. Second, we exploit randomization of the neighboring school's treatment as exogenous variation in switching costs.

6.1 Sorting by teacher information

A 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 baseline 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] [under the performance raise contract]."

We find suggestive evidence that teachers who were provided information about their performance exhibit stronger sorting effects. The correlation between choosing performance pay and teachers' value-added increases by 50% for those assigned to the information treatment versus no information, as we see in figure 6 and table A6. With the information treatment, the difference in value-added percentile for teachers selecting performance pay versus flat pay is 9.8 percentile versus 6.8 percentile for the control group (significant at the 10% level). This change comes from both high value-added teachers being more likely to opt into performance pay and low value-added teachers being more likely to select flat pay when under the information treatment.

6.2 Sorting by switching costs

The extent of positive sorting may also depend on how strong employees preferences are for wages versus amenities, such as location or school characteristics. 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 closest school. Most schools operate on a larger campus, which contains multiple schools (primary school, middle school, high school). 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 A12, 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:

$$\begin{aligned} VA_{i,t-1} = \beta_0 \quad & +\beta_1 WorkatPP_i + \beta_2 Post_i + \beta_3 WorkatPP_i \cdot Post_i + \beta_4 OppTreat_i \\ & +\beta_5 OppTreat_i \cdot Post + \beta_6 OppTreat_i \cdot WorkatPP_i \\ & +\beta_7 OppTreat_i \cdot WorkatPP_i \cdot Post + \chi_j + \epsilon_i \end{aligned} \tag{10}$$

This is similar to eq. 8 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 but the difference is not statistically significant (p-value = 0.15). Figure 7 and table A7 presents these results. Column 1 shows the extent of positive sorting for teachers whose closest school has the same treatment, and column 2 shows the results when the closest school has the opposite treatment. The magnitude of positive sorting is about four times larger (0.040 SD versus 0.009 SD).

Another approach to test whether switching costs dampen the extent of positive sorting is to compare the contract choice versus the job choice results. 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 results from these two analyses (table 6), we see substantial differences in the extent of positive sorting (0.074 SD versus 0.033 SD). This gives us a lower and upper bound on the sorting effect in the long term.

7 Testing for negative sorting

7.1 Does performance pay attract bad actors?

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 two categories: i). teaching pedagogy (using classroom observation data), ii). student socio-emotional development (using a student survey).

First, we do not find that teachers who prefer performance pay are more likely to engage in distortionary teaching practices. In fact, they are significantly *less* likely to exhibit these behaviors than teachers who did not want performance pay. Figure 8, panel A, and appendix table A8 presents the treatment effects of objective performance pay along each teacher behavior, measured via classroom observation. Teachers are scored 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, 2024), we find that objective performance pay results in a more negative classroom climate (more yelling, stricter discipline), more teacher-led time (more lecturing), 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. We can reject equality of treatment effects between teachers who chose performance pay versus flat pay at the 1% level.

Second, we do *not* find that teachers who prefer performance pay ignore other areas of student development in order to maximize their pay. As with the teacher behavior results, teachers who prefer performance pay are less likely to exhibit distortionary behavior when assigned to performance pay contracts than those who wanted flat pay. Figure 8, panel B and appendix table A9 present results. At endline, we measure student satisfaction and socio-emotional development along five dimensions from an endline survey with students (survey items and sources shown in appendix table D.2). 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. We can reject equality of treatment effects between teachers who chose performance pay versus flat pay at the 1% level.

7.2 Does performance pay push out good actors?

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 extracurriculars).[21] Figure A7 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.

8 Discussion & Conclusion

8.1 Connecting Previous Evidence

Our conceptual framework gives us a unifying theory which can help to square a number of previous findings studying the relationship between performance pay and selection. Table A10 describes results from the most prominent other papers which have estimated selection versus effort responses to performance pay in any sector. Across these studies, selection effects have varied substantially, but so too has the extent of private information and switching costs agents would bear to receive their preferred contract.

While the other studies do not explicitly report measures of private information or switching costs, given the setting we can make a best guess. For example, in Dohmen and Falk (2011) lab participants are solving math problems and they first solve a number of similar problems under a piece rate scheme and then are asked whether they'd like to solve more under flat pay or piece rate. Here the agent is well informed about the expected payoffs (high information) and they are able to directly pick their contract (zero switching cost). Alternatively in Leaver et al. (2021), the sample is new entrants to the teaching profession choosing between jobs in certain districts/subjects offering performance pay contract based on value added, attendance and lesson plan preparation versus those in other offering play pay. The individuals have not had experience teaching so likely have minimal private information (given what we find among novice teachers in our sample) and would either have to move districts or re-license in a new subject to receive their preferred contract (high switching cost).

[21] Survey item description and sources are presented in appendix table D.1. 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.

Figure A14 plots the relationship between the extent of private information, switching costs and selection effects in these studies. Consistent with our findings, papers which focused on settings where agents have more private information and lower switching costs found larger selection effects.

8.2 Policy Counterfactuals

While the results of this experiment allow us to measure the extent of asymmetric information between employees and employers, the estimates likely do not reflect the general equilibrium sorting effects if the entire teaching profession switched to performance pay. The introduction of a longer-term, district or profession-wide performance pay system for teachers would have several effects which are different than our experiment: i). teachers and potential teachers would need to sort across district or occupation rather than just across employment location to receive their preferred contract (increasing up-front switching costs), ii). payoff for switching into the preferred contract would accrue each year the employee is under their preferred contract (increasing long term switching benefits).

To estimate the combined effect of these two additional forces on teacher sorting, we augment the model from section 2.1 by allowing workers to choose each year between many jobs (both in and out of the teaching sector), adding a cost to change districts/occupation, and allowing heterogeneity in private information by worker experience. The key moments and parameters come from experimental results, administrative data and survey data (table A11). The model and estimation is described in detail in appendix B.

We find that the extent of positive sorting depends heavily on the length and type of policy (appendix figure A15). For a long term policy paying teachers 20% of their salary in the form of performance pay, introduced at a district level, the effect on test scores is 0.10-0.21 SD.[22] This is 1.5-3.1x the effort response effect of performance pay, suggesting the sorting component could be an important margin. The effects are smaller but still substantial, if the policy is introduced at the occupation level (0.09 SD-0.17 SD) or at the district level for a shorter term, 10 year, policy (0.07-0.10 SD).

8.3 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 who have higher baseline value-added and have larger behavioral responses

[22] The range of values includes optimistic and pessimistic values for parameter values for the non-teaching population which we had to extrapolate to using our sample.

significantly prefer performance pay. Teachers with higher baseline value-added are also more likely to switch into performance pay schools once the contracts are in place. 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.

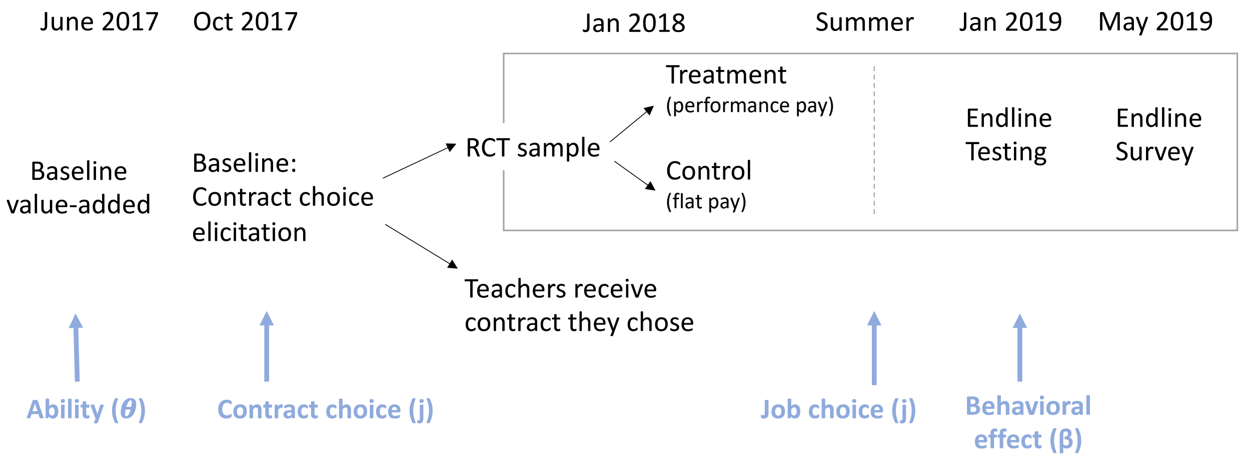
To understand what the effects of different policies would be on the extent of sorting, we use additional exogenous variation to test the effect of increasing teachers private information and lowering the switching cost to access their preferred contract. We find teachers are responsive to both of these 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 policy at the occupation level are 1.3-2.4x the effects of a one year policy due to sorting.

Two limitations of the study are the inability to look at long run effects directly in the experimental sample and needing to extrapolate our estimates to other populations (e.g. pre-entry teachers). Understanding the features of this population is an important area for further work. In addition, understanding what sectors high quality potential teachers are drawn from is crucial, 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 the information employees have to help improve the quality and match of their workforce. 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 estimating the effect of performance pay on the current population of teachers may underestimate the effects in the long run by ignoring the sorting response to performance pay.

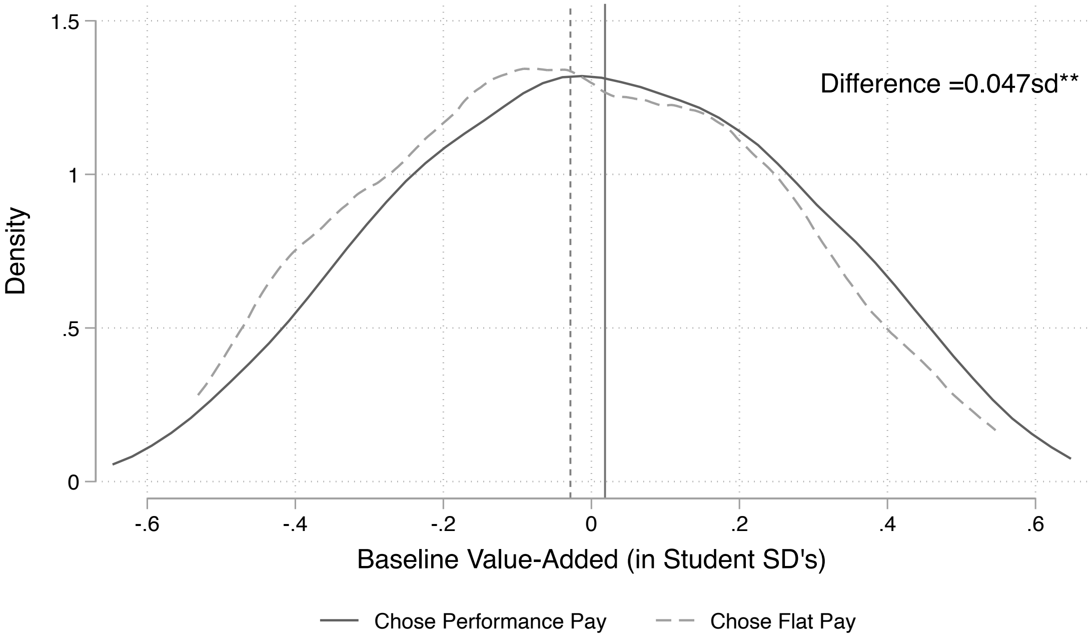
9 Figures

Figure 1: Experiment Timeline



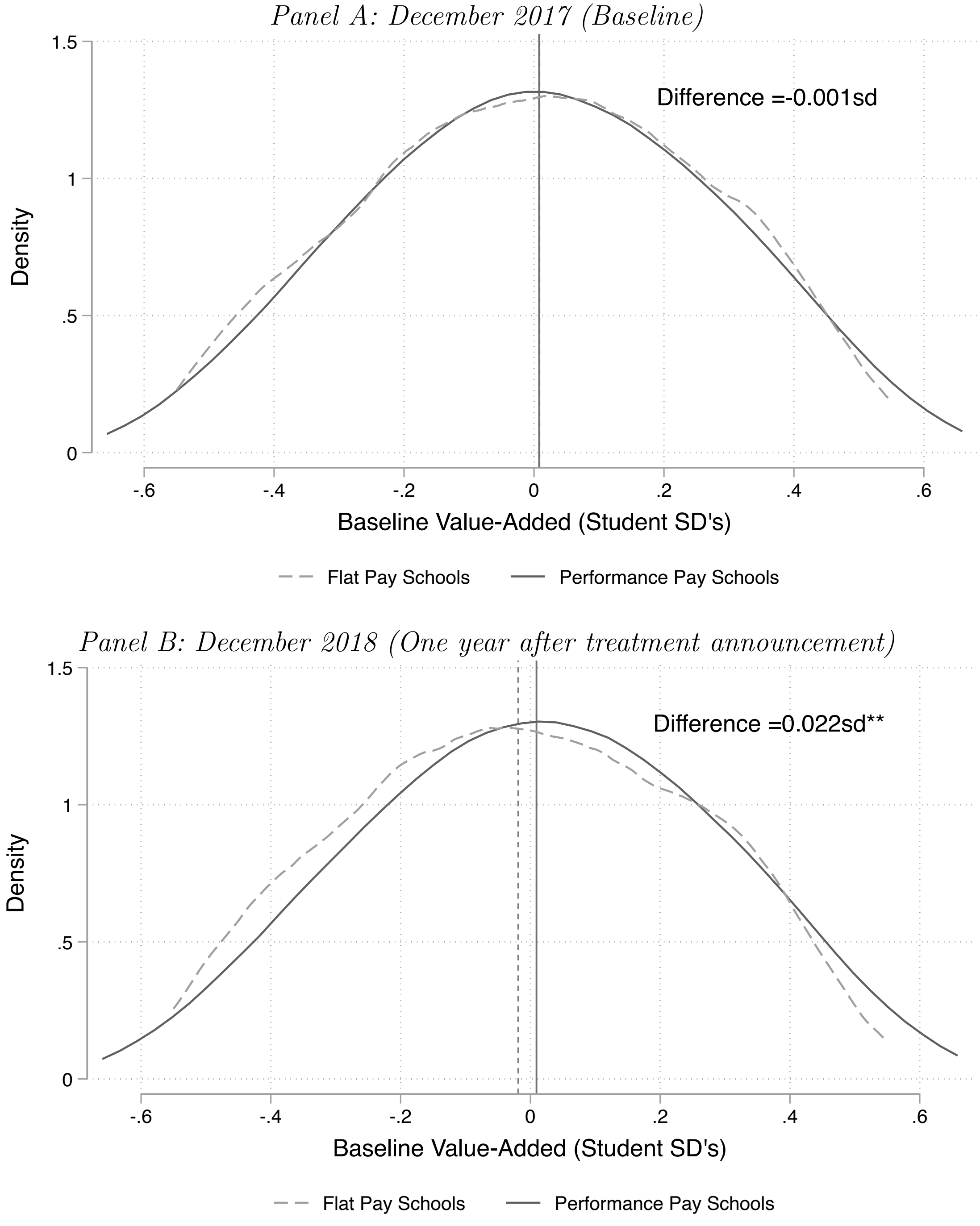
Notes: The figure presents the experimental timeline from June 2017 through May 2019. Our measure of ability comes from the calculation of teacher value-added in June 2017 prior to the introduction of the treatments. Our measure of the behavioral effect of performance pay comes from comparing the treatment and control sample in January 2019, a year after the introduction of the new contracts. We measure teacher job choice 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 2: Distribution of Baseline Value-Added by Contract Choice



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

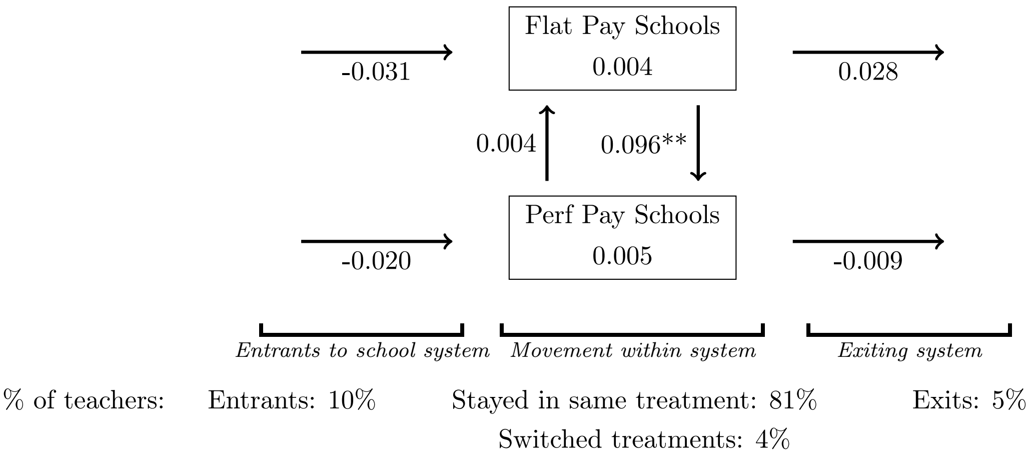
Figure 3: Distribution of Teacher Baseline Value-Added by School and Year


Panel A: December 2017 (Baseline)

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

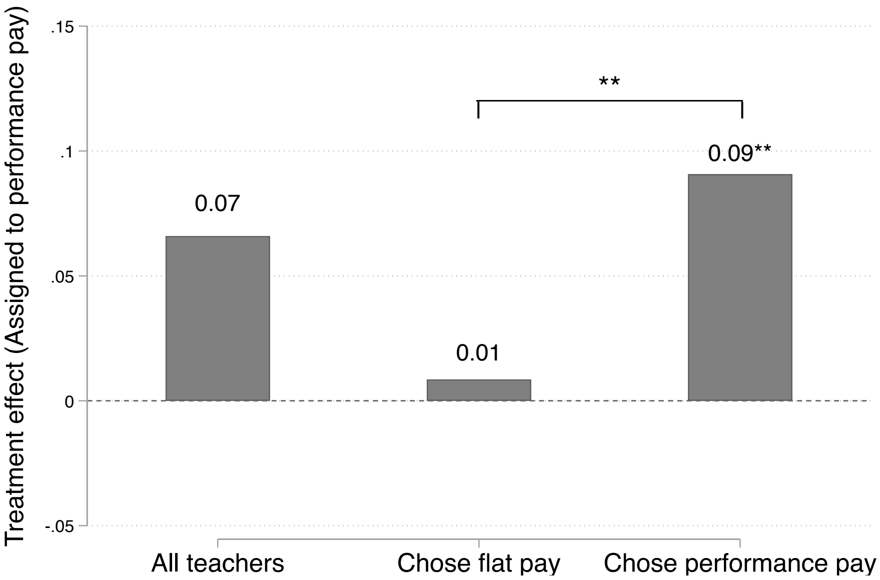
Notes: These figures plot the distribution of baseline teacher value-added for teachers in performance pay versus flat pay schools. Panel A provides the distribution in December 2017 (one month before the treatments are announced). Panel B provides the distribution in December 2018 (11 months after the treatments are announced). Teacher employment data comes from school administrative records. Value-added is calculated using three 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 4: Average teacher value-added by employment flows



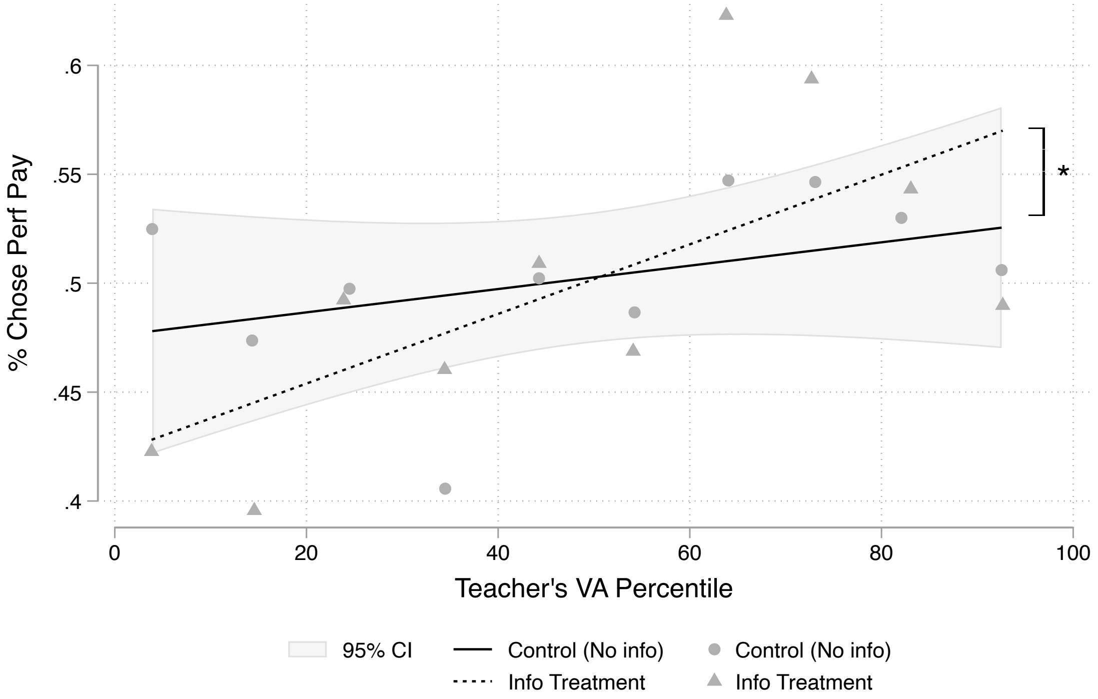
Notes: This figure shows the average baseline value-added for new entrants into this school system, teachers staying at their current school or switching to a different school within the system and those exiting the system. The top half of the diagram presents this for those in flat pay schools and bottom half shows this for performance pay schools. Finally, the figures at the very bottom show what percentage of the total teacher population fall into each category. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Figure 5: Treatment Effect by Contract Choice



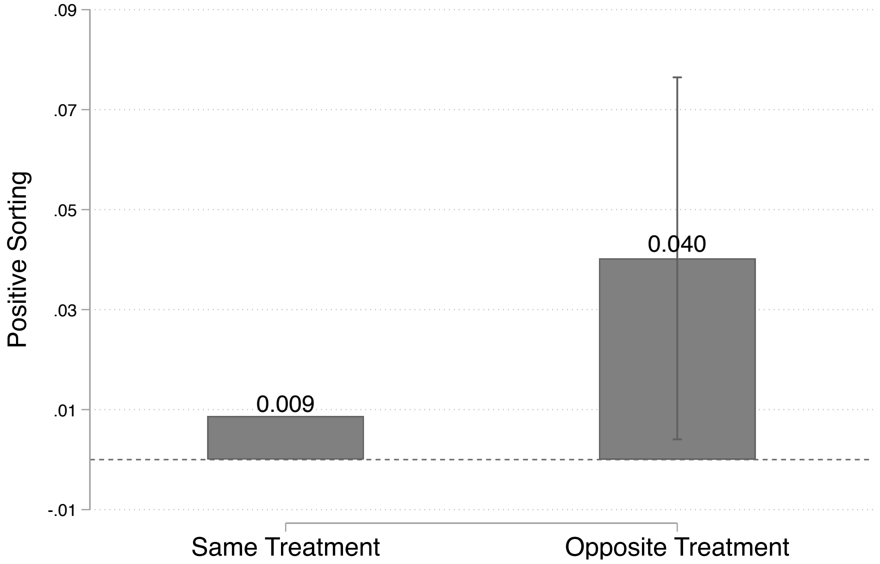
Notes: This figure presents the effect of being assigned performance pay relative to flat pay on endline test scores. The first bar presents the effects for all teachers. The second bar presents the treatment effects for teachers who stated in the baseline contract choice exercise that they wanted a flat pay contract and the third bar those who wanted a performance pay contract. Standard errors are clustered at the school level. $^{*}p < 0.10, ^{**}p < 0.05, ^{***}p < 0.01$.

Figure 6: Contract Choice by Teacher Value-Added and Treatment



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

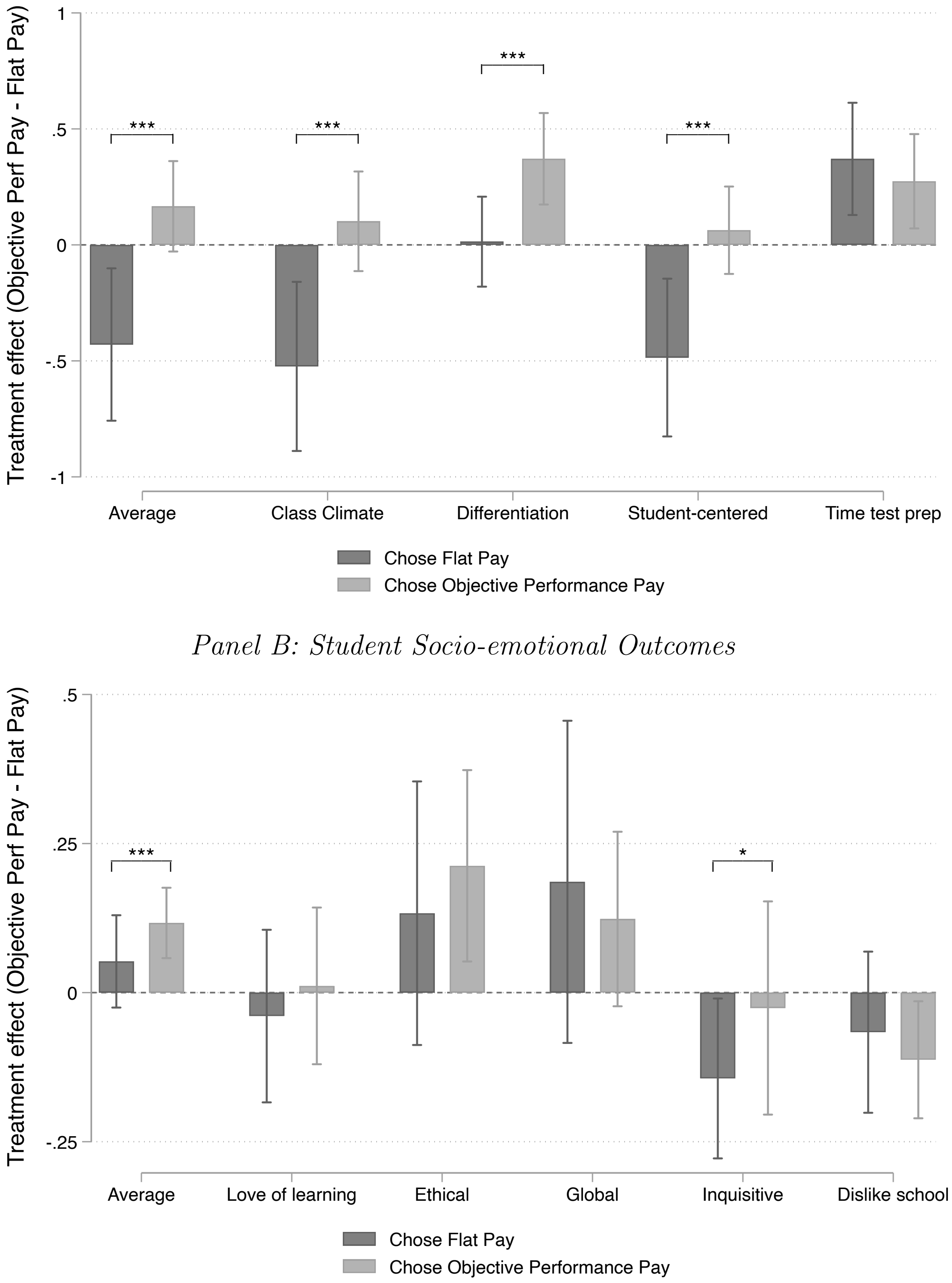
Figure 7: Positive Sorting by Closest School's Treatment



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

Figure 8: Treatment Effects on Non-Test Score Outcomes by Contract Choice

Panel A: Teaching Quality



Panel B: Student Socio-emotional Outcomes

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 (panel A) and student surveys (panel B) . Standard errors are clustered at the school level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

10 Tables

Table 1: Performance Raise Categories

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%

Notes: This table shows the performance categories and corresponding raise value used in the treatment schools.

Table 2: Teacher Value-Added by Contract Choice

	Teacher Baseline Value-Added (in Student SDs)					
	(1)	(2)	(3)	(4)	(5)	(6)
Chose Performance Pay	0.0468** (0.0185)	0.0424** (0.0186)	0.0485** (0.0207)	0.0455** (0.0207)	0.0452** (0.0218)	0.0392* (0.0221)
Principal Rating of Teacher		0.0203* (0.0104)		0.0208** (0.0104)		0.0200* (0.0105)
Observations	2568	2568	1284	1284	1284	1284
Performance Metric	All	All	Objective	Objective	Subjective	Subjective
Control Mean	-0.0283	-0.0283	-0.0283	-0.0283	-0.0284	-0.0284
Control SD	0.347	0.347	0.349	0.349	0.345	0.345

Notes: This table presents the relationship between teacher contract choice and baseline value-added. *Teacher Baseline Value-Added* is a measure of teacher value-added using test score data from the three 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. *Principal Rating of Teacher* is the principal's evaluation score of the teacher (as a standardized z-score). Columns (1) and (2) include two observations per teacher: the choice between objective (value-added based) performance pay and flat pay and the choice between subjective (principal evaluation based) performance pay and flat pay. Columns (3)-(6) restrict to just the choice between objective and flat or just subjective and flat. Standard errors are clustered at the teacher level. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table 3: Teacher Quality by School

	Teacher Baseline Value-Added (in Student SDs)			
	(1)	(2)	(3)	(4)
Performance Pay Schools	-0.0160	-0.0143	0.00178	0.00115
	(0.0188)	(0.0189)	(0.0198)	(0.0197)
Post	-0.0191*	-0.0194*	-0.0195*	-0.0205*
	(0.0107)	(0.0108)	(0.0108)	(0.0107)
Performance Pay Schools*Post	0.0222**	0.0225**	0.0231**	0.0237**
	(0.0113)	(0.0113)	(0.0113)	(0.0112)
Principal Rating of Teacher				0.0197***
				(0.00707)
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	0.0190
Control SD	0.327	0.327	0.327	0.327
Clusters	243	243	243	243
Observations	6991	6991	6991	6991

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 three 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. 8). Columns (2)-(4) add additional controls. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table 4: 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	0.00785	0.0630	0.00160
	(0.0397)	(0.0408)	(0.0511)	(0.0510)	(0.0421)	(0.0551)
Chose Perf Pay* Assigned Perf Pay Treat			0.0822**	0.0819**		0.0882**
			(0.0406)	(0.0406)		(0.0440)
Principal Rating of Teacher				0.00324		
				(0.00986)		
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 Controls	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 worked at a school assigned to performance pay at the time of the treatment announcement. *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 principal's evaluation score of the teacher (as a standardized z-score). 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 baseline value-added, respectively. Column (6) combines the two, and column (4) controls for principal's rating of the teacher. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table 5: Treatment Effect by Contract Choice without Controls

	Endline Test (z-score)			
	(1)	(2)	(3)	(4)
Assigned Perf Pay Treat	-0.0342	-0.0348	0.0213	-0.0468
	(0.0437)	(0.0436)	(0.0405)	(0.0472)
Chose Perf Pay* Assigned Perf Pay Treat	0.0886**	0.0886**		0.0982**
	(0.0388)	(0.0388)		(0.0418)
Principal Rating of Teacher		0.00315		
		(0.0107)		
Baseline Value-Added*Assigned Perf Pay Treat			-0.0783	-0.0921
			(0.134)	(0.135)
Control Mean	7.94e-10	7.94e-10	-0.00223	-0.00223
Control SD	1.000	1.000	0.997	0.997
Clusters	114	114	109	109
Observations	144009	144009	126989	126989
Randomization Strata Fixed Effects	Yes	Yes	Yes	Yes
Baseline Controls	No	No	No	No

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 worked at a school assigned to performance pay at the time of the treatment announcement. *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 principal's evaluation score of the teacher (as a standardized z-score). 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 baseline value-added, respectively. Column (6) combines the two, and column (4) controls for principal's rating of the teacher. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table 6: Total Effect of Performance Pay on Student Test Scores

Research design:	Contract Choice		Job Choice	
	Total effect (1)	% of total (2)	Total effect (3)	% of total (4)
1. Total Sorting effect:	0.074	53%	0.033	33%
1a. Sorting on ability	0.049	35%	0.022	22%
1b. Sorting on behavioral effect	0.025	18%	0.011	11%
2. Behavioral effect:	0.066	47%	0.066	67%
Total effect of performance pay	0.140	100%	0.099	100%

Notes: This table shows the total effect of performance pay on student test scores in terms of standard deviations. The rows show the decomposition of the sorting effect versus the behavioral response to performance pay. Column 1 shows the estimates of these effects using the contract choice research design and column 3 uses the actual observed movement across schools in second year. Column 2 and 4 show what percentage of the total effect of performance pay comes from sorting versus behavioral effects.

Table 7: Principal Beliefs about Teacher Quality

Panel A: Beliefs by outcome type	Principal Belief (z-score)			
	(1)	(2)	(3)	(4)
Teacher outcome type:	Attendance	Discipline	Pedagogy	VA
Teacher Outcome (z-score)	0.192***	0.231**	0.136	-0.0484
	(0.0503)	(0.104)	(0.125)	(0.0831)
Observations	250	143	143	166

Panel B: Beliefs by Principal Characteristic	Principal Belief (z-score)			
Teacher outcome type:	All	All	All	All
Teacher Outcome (z-score)	0.166***	0.0579	0.184***	0.150***
	(0.0434)	(0.0680)	(0.0482)	(0.0383)
Observation treatment		-0.0433		
		(0.0900)		
Teacher Outcome*Observation treatment		0.195*		
		(0.1000)		
Overlap > 2 years with teacher			0.165*	0.111
			(0.0850)	(0.0977)
Teacher Outcome*Overlap > 2 years			-0.182**	-0.155**
			(0.0806)	(0.0700)
Dep. Var. Mean	-0.0351	-0.0351	-0.0351	-0.0351
Dep. Var. SD	1.003	1.003	1.003	1.003
Observations	702	594	702	702
Principal Fixed Effects	No	No	No	Yes

Notes: This table presents the relationship between teacher's actual performance 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 (pedagogy) and value-added. *Principal beliefs* are from principal endline survey data. *Teacher outcome* is the teacher's actual performance on each outcome from administrative data (for *attendance* and *value-added*) and classroom observation data (for *discipline* and *pedagogy*). Panel A presents the relationship between actual performance and principal beliefs by outcome type. Panel B pools across all four outcomes and looks at the interaction with principal characteristics. *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. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

References

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

Andrabi, Tahir and Christina Brown, "Subjective and Objective Incentives and Employee Productivity," *Working Paper*, July 2024, p. 52.

ASER, *Annual Status of Education Report Pakistan* 2019.

Ashraf, Nava, 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*, May 2020, *110* (5), 1355–1394.

Azam, Mehtabul and Geeta Gandhi Kingdon, "Assessing teacher quality in India," *Journal of Development Economics*, November 2015, *117*, 74–83.

Barlevy, Gadi and Derek Neal, "Pay for Percentile," *American Economic Review*, August 2012, *102* (5), 1805–1831.

Bau, Natalie and Jishnu Das, "Teacher Value Added in a Low-Income Country," *American Economic Journal: Economic Policy*, February 2020, *12* (1), 62–96.

Biasi, Barbara, "The Labor Market for Teachers under Different Pay Schemes," *American Economic Journal: Economic Policy*, August 2021, *13* (3), 63–102.

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

Burrell, David L, "Relationships Among Teachers' Efficacy, Teachers' Locus-of-control, and Student Achievement," 1994.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, September 2014, *104* (9), 2633–2679.

Dee, Thomas S. and James Wyckoff, "Incentives, Selection, and Teacher Performance: Evidence from IMPACT: Incentives, Selection, and Teacher Performance," *Journal of Policy Analysis and Management*, March 2015, *34* (2), 267–297.

Dohmen, Thomas and Armin Falk, "Performance Pay and Multidimensional Sorting: Productivity, Preferences, and Gender," *American Economic Review*, April 2011, *101* (2), 556–590.

Finkelstein, Amy and Kathleen McGarry, "Multiple Dimensions of Private Information: Evidence from the Long-Term Care Insurance Market," *American Economic Review*, August 2006, *96* (4), 938–958.

Fryer, Roland G., "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools," *Journal of Labor Economics*, April 2013, *31* (2), 373–407.

Goodman, Sarena F. 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*, April 2013, *31* (2), 409–420.

Greenwald, Bruce C., "Adverse Selection in the Labour Market," *The Review of Economic Studies*, July 1986, *53* (3), 325.

Jackson, C. Kirabo, "What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes," *Journal of Political Economy*, October 2018, *126* (5), 2072–2107. Publisher: The University of Chicago Press.

Johnston, Andrew C., "Teacher Preferences, Working Conditions, and Compensation Structure," *SSRN Electronic Journal*, 2020.

Kane, Thomas and Douglas Staiger, "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," Technical Report w14607, National Bureau of Economic Research, Cambridge, MA December 2008.

Kantarevic, Jasmin and Boris Kralj, "Physician Payment Contracts in the Presence of Moral Hazard and Adverse Selection: The Theory and Its Application in Ontario," *25* (10), 1326–1340.

Karlan, Dean and Jonathan Zinman, "Observing Unobservables: Identifying Information Asymmetries With a Consumer Credit Field Experiment," *Econometrica*, 2009, *77* (6), 1993–2008.

Lavy, Victor, "Performance Pay and Teachers' Effort, Productivity, and Grading Ethics," *American Economic Review*, December 2009, *99* (5), 1979–2011.

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

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

Leaver, Clare, Owen Ozier, Pieter Serneels, and Andrew Zeitlin, "Recruitment, Effort, and Retention Effects of Performance Contracts for Civil Servants: Experimental Evidence from Rwandan Primary Schools," *American Economic Review*, July 2021, *111* (7), 2213–2246.

Muralidharan, Karthik and Venkatesh Sundararaman, "Teacher Performance Pay: Experimental Evidence from India," *Journal of Political Economy*, February 2011, *119* (1), 39–77.

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

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

Rockoff, Jonah E. and Cecilia Speroni, "Subjective and Objective Evaluations of Teacher Effectiveness," *The American Economic Review*, 2010, *100* (2,), 261–266.

Rose, Evan K, Jonathan Schellenberg, and Yotam Shem-Tov, "The Effects of Teacher Quality on Adult Criminal Justice Contact," *NBER Working Paper*, July 2022, *No. 30274.*

Rothstein, Jesse, "Teacher Quality Policy When Supply Matters," *American Economic Review*, January 2015, *105* (1), 100–130.

Roy, A D, "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, June 1951, *3* (2), 135–146.

Springer, Matthew G, 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.

Staiger, Douglas O and Jonah E Rockoff, "Searching for Effective Teachers with Imperfect Information," *Journal of Economic Perspectives*, August 2010, *24* (3), 97–118.

The World Bank , *Enterprise Surveys* 2019.

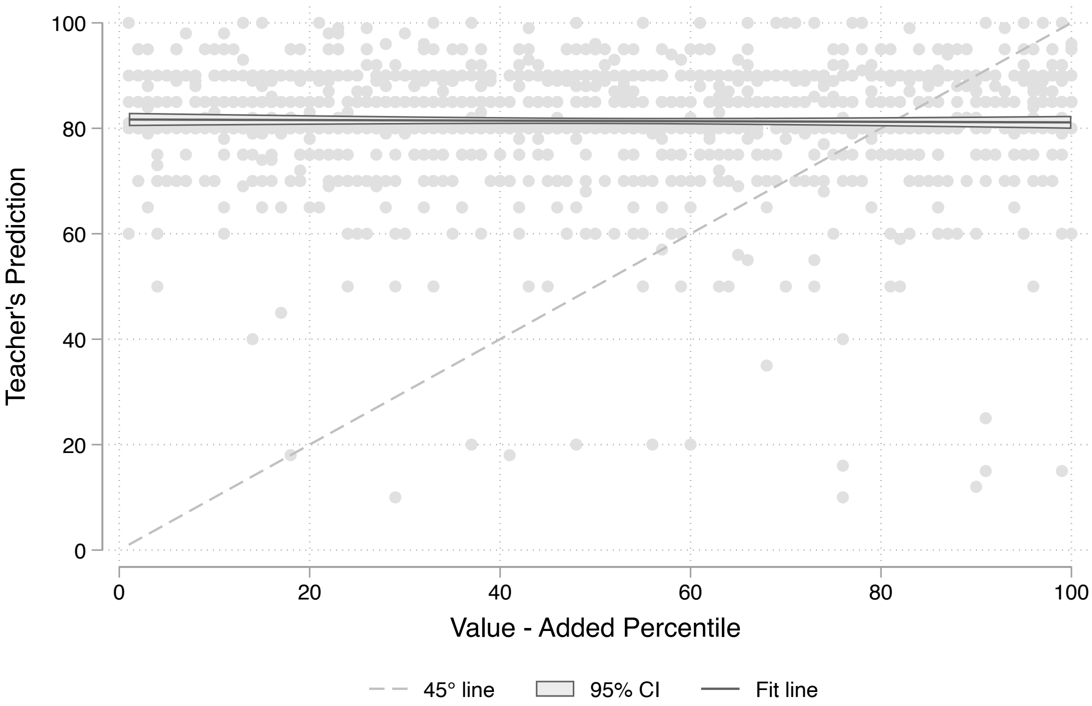
The World Bank Group, *Systems Approach for Better Education Results (SABER)* 2018.

_ , *World Development Report 2018: Learning to Realize Education's Promise*, Washington, DC: World Bank, 2018.

Online Appendix

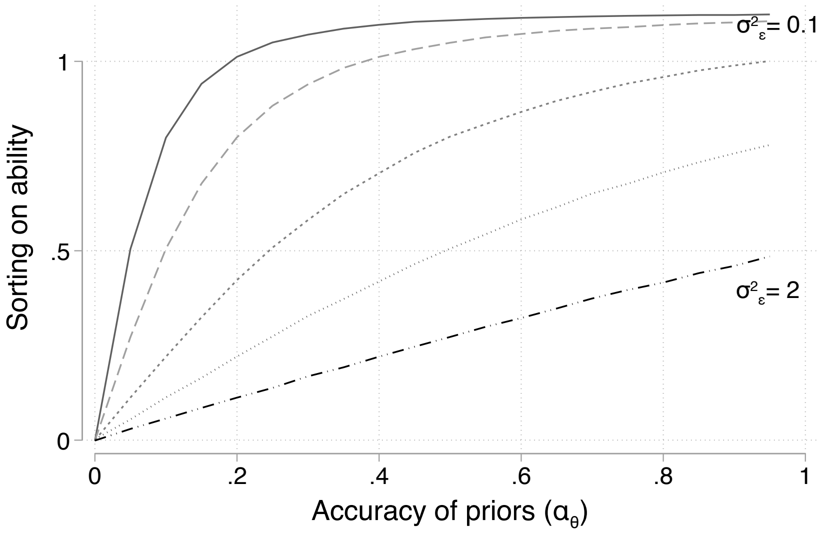
A Appendix Figures and Tables

Figure A1: Teacher Beliefs about Value-Added



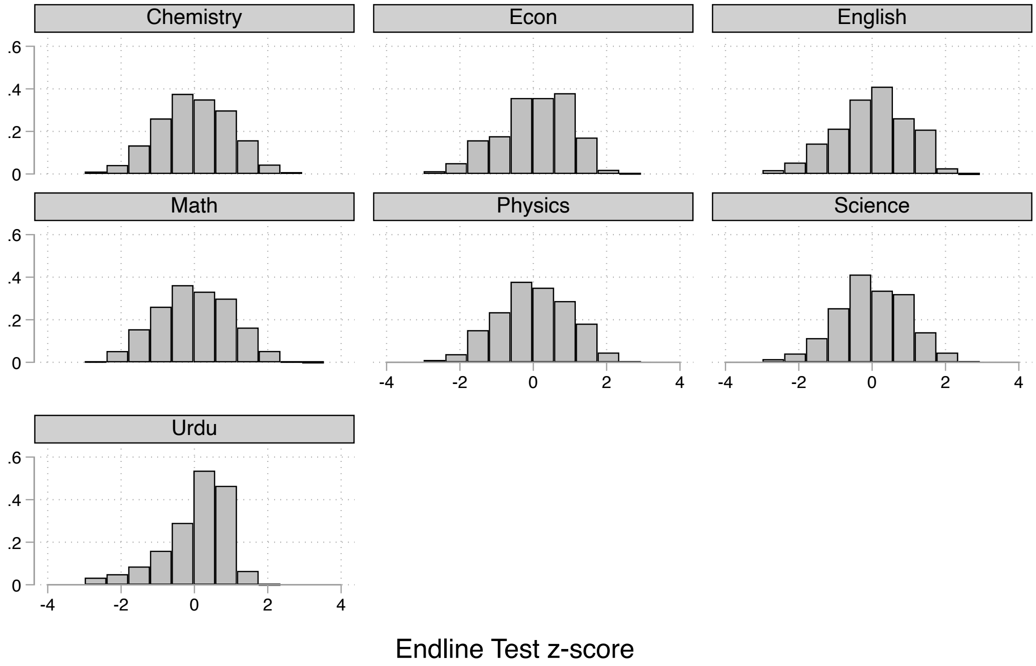
Notes: This figure shows the relationship between teacher's stated prediction of their value-added percentile relative to their actual value-added percentile. Teacher predictions come from the baseline survey and actual value-added is from administrative test data. Each point is a teacher's response and the solid line and shaded area show the 95% confidence interval.

Figure A2: Extent of Sorting by Model Parameters



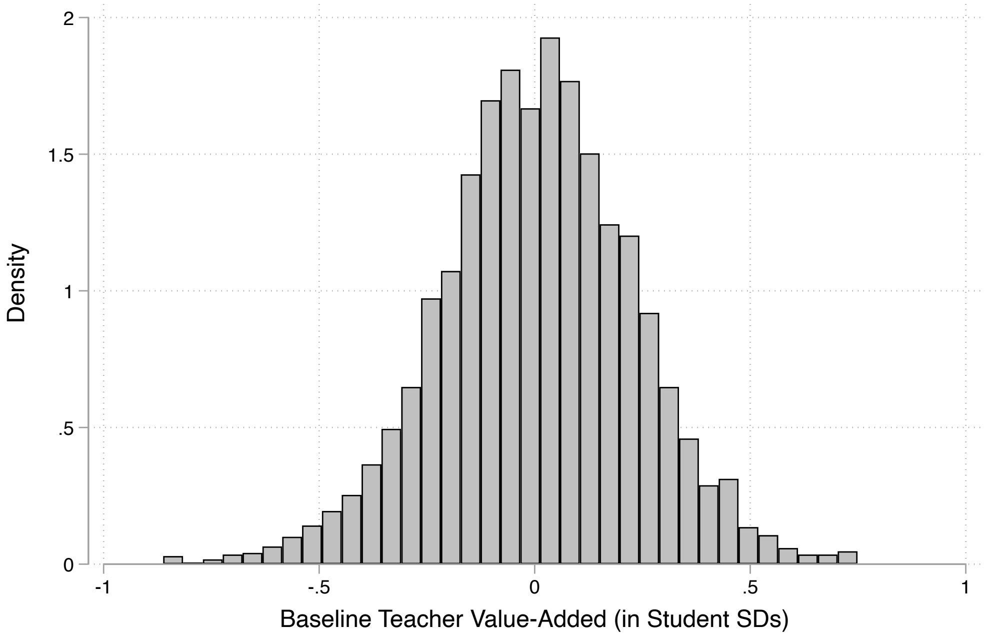
Notes: This figure simulates the extent of sorting on ability based on the model for varying parameter values. Each line represents the level of sorting on ability we would expect to see for varying levels of teacher accuracy about their ability (from no private information, $\alpha_\theta = 0$, to full information, $\alpha_\theta = 1$). Across the different lines we impute different values of the variance in non-wage utility, ranging from a very low variance ($\sigma^2_\epsilon = 0.1$) for the solid line up to a very high variance ($\sigma^2_\epsilon = 0.1$) for the large dashed line.

Figure A3: Distribution of Endline Test Scores



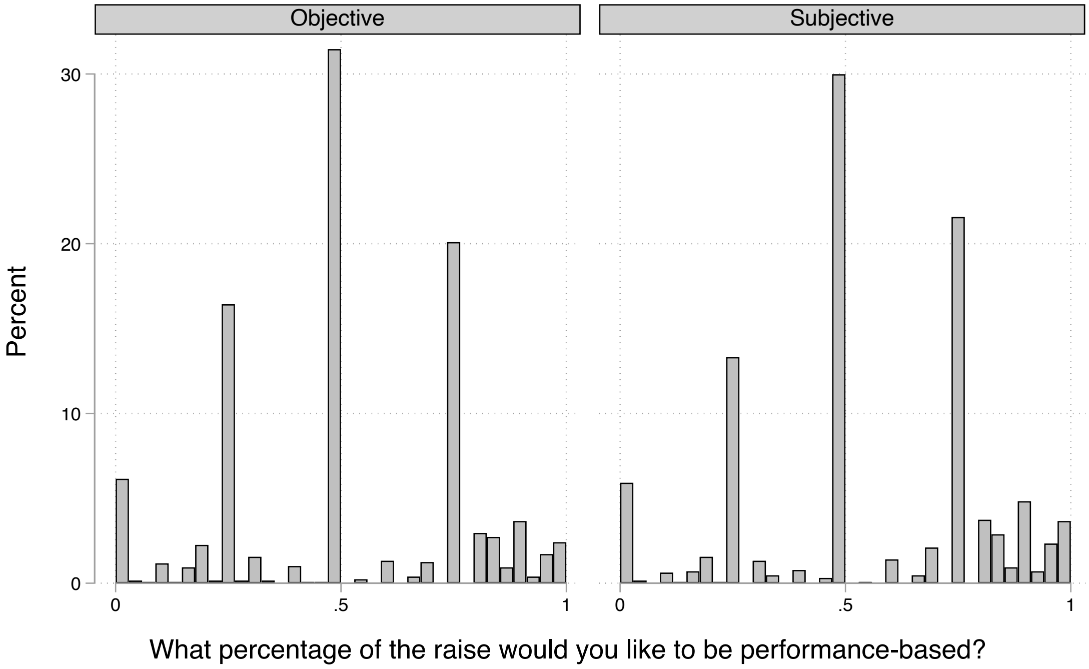
Notes: This figure presents the standardized distribution of student scores across each exam subject administered at endline for 48,148 students. The endline test was conducted in January 2019 across grades 4-13 in English, Urdu, Math, Science and Economics. In grades 9-13, students took the science exam in the class they were currently enrolled, either Chemistry or Physics.

Figure A4: Distribution of Teacher Value-Added at Baseline



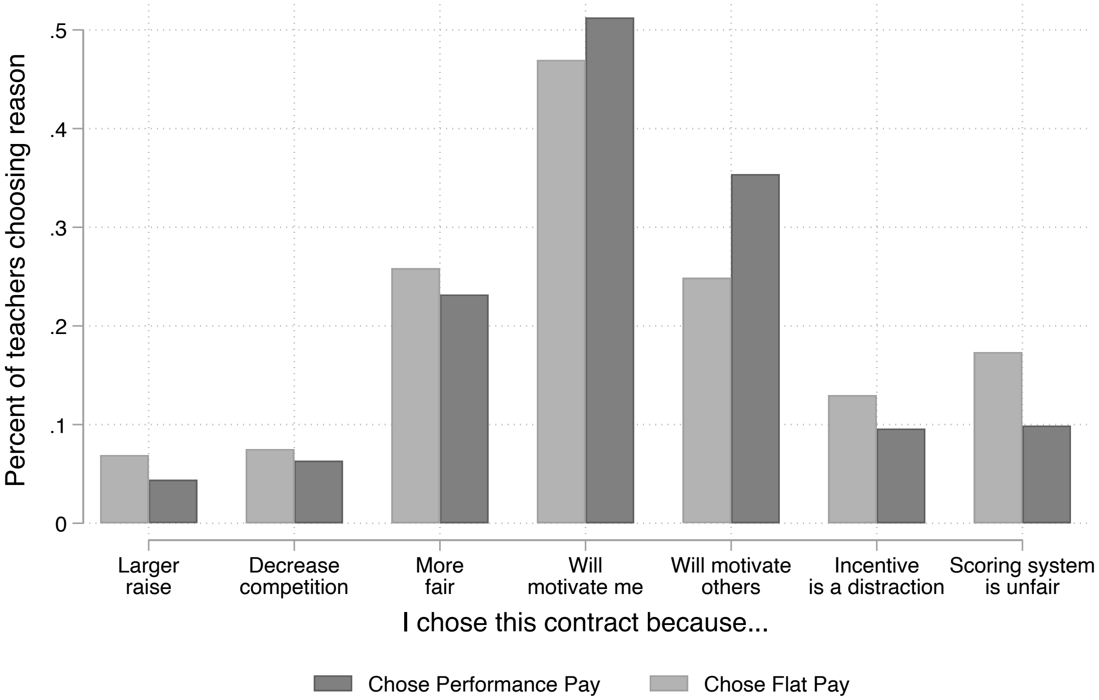
Notes: This figure presents the distribution of teacher value-added for 3,687 teachers in the school system at baseline. Teacher value-added is calculated using administrative test score data from June 2015, 2016 and 2017 (the three years prior to the intervention). Estimates are calculated following Kane and Staiger (2008), using an empirical Bayes approach.

Figure A5: Distribution of contract choice by performance metric



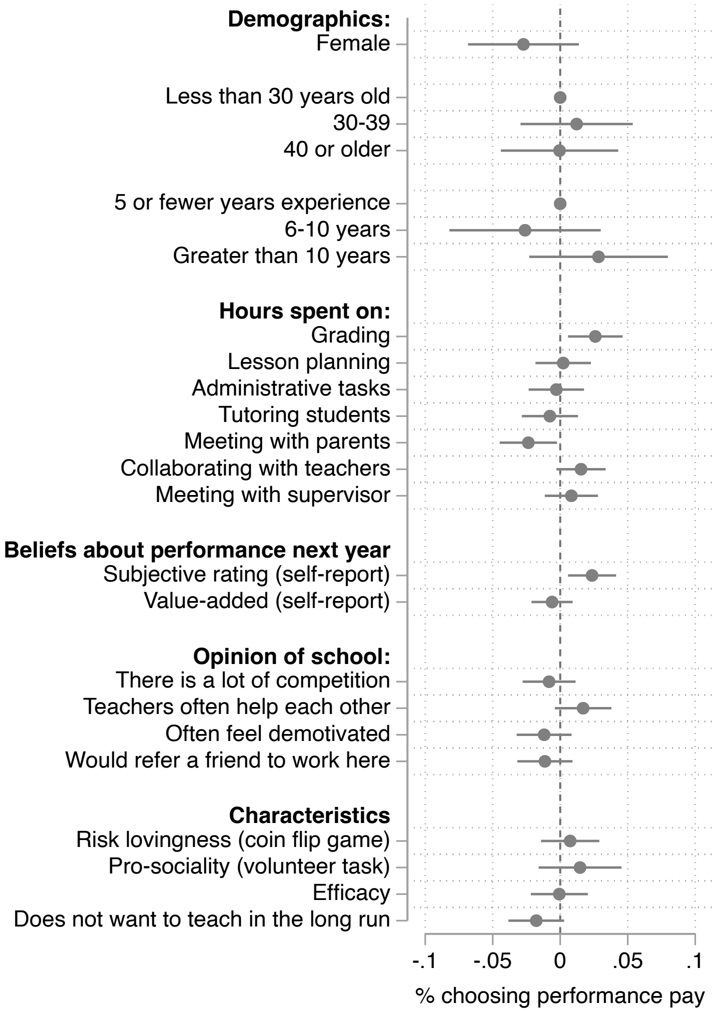
Notes: These figures plot teachers' survey response to the contract choice question for all 1284 teachers who completed the baseline survey. We ask teachers: *We can think of a raise as being a combination of two parts: the "flat" part that everyone gets regardless of their [subjective/objective] score 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 A6: Teachers stated reasons for selecting performance pay or flat pay contract



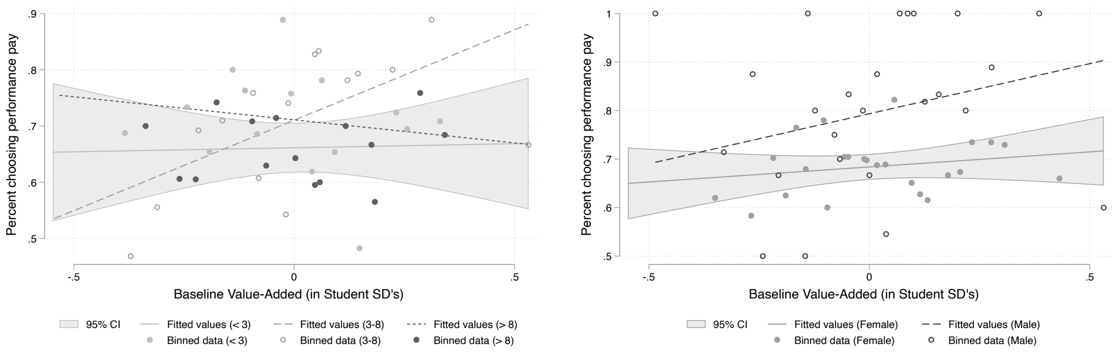
Notes: This figure plots teachers' responses to the question: *Why did you select this contract?* for all 1284 teachers who completed the baseline survey. The graph shows the percent of teachers that selected each reason. Teachers are allows to select multiple reasons, if applicable. The light gray bars plot responses for teachers who chose a flat pay contract. The dark gray bars plot responses for teachers who chose performance pay contracts.

Figure A7: Predictors of contract choice



Notes: This figure presents coefficients and 95% confidence intervals of bivariate regressions of teacher's contract choice on teacher demographics, time use, 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, time use beliefs and contract choice come from the baseline survey. Standard errors are clustered at the teacher level.

Figure A8: Relationship between Value-Added and Contract Choice

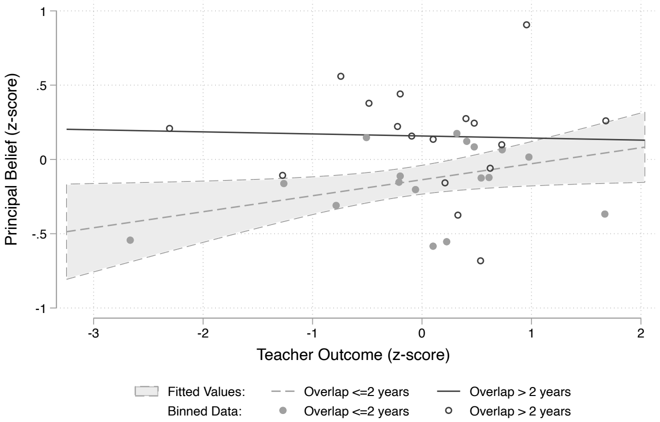


Panel A: By Teacher Experience

Panel B: By Teacher Gender

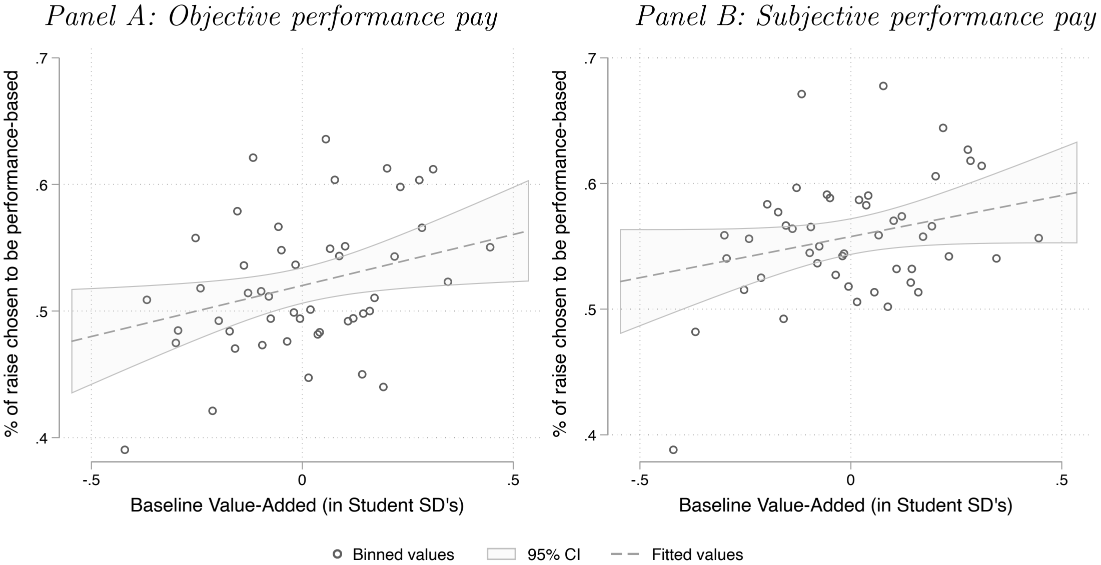
Notes: These figures plot the relationship between teacher quality as measured by baseline value-added and teacher's contract choice by teacher characteristic for all 1284 teachers who completed the baseline survey. The graph plots binned values of *Teacher Baseline Value-Added* by the percent of teachers in that bin that chose performance pay. The shaded area shows the 95% confidence interval. Panel A present results by teacher's number of years of experience. Panel B presents results by teacher gender. Choice data comes from the contract choice exercise conducted in October 2017. Value-added is calculated using three years of administrative data prior to the start of the intervention.

Figure A9: Principal Beliefs about Teacher Outcome by Years of Relationship



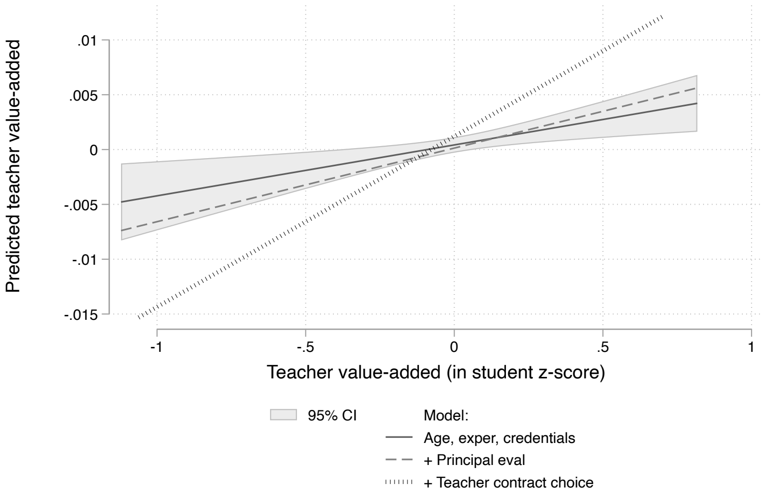
Notes: This figure presents principals' beliefs about teacher quality versus their actual performance. *Principal beliefs* are measured in z-scores and come from endline surveys with principals. *Teacher outcome* is the the teacher's z-score in each of four outcomes: value-added, attendance, behavioral management and use of analysis/inquiry. The shaded area shows the 95% confidence interval. Baseline value-added is calculated using three years of administrative data prior to the start of the intervention. 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). Standard errors are clustered at the teacher level.

Figure A10: Contract Choice by Teacher Value-Added



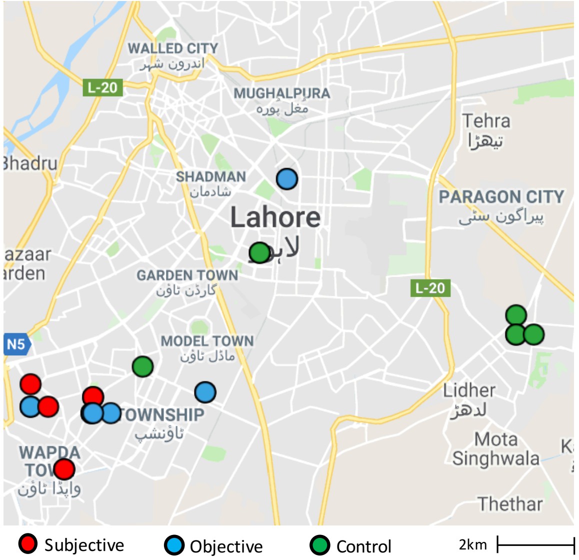
Notes: The figure shows the relationship between teacher's contract choice and value added. Panel A is for the choice between flat pay versus objective performance pay (where teachers could select a value from 0 (fully flat) to 100 (fully performance-based)). Panel B is the choice between flat and subjective performance pay. Contract choice data come from the baseline survey conducted in October 2017. Value-added is calculated using three years of administrative data prior to the start of the intervention.

Figure A11: Predicting Teacher Value-Added



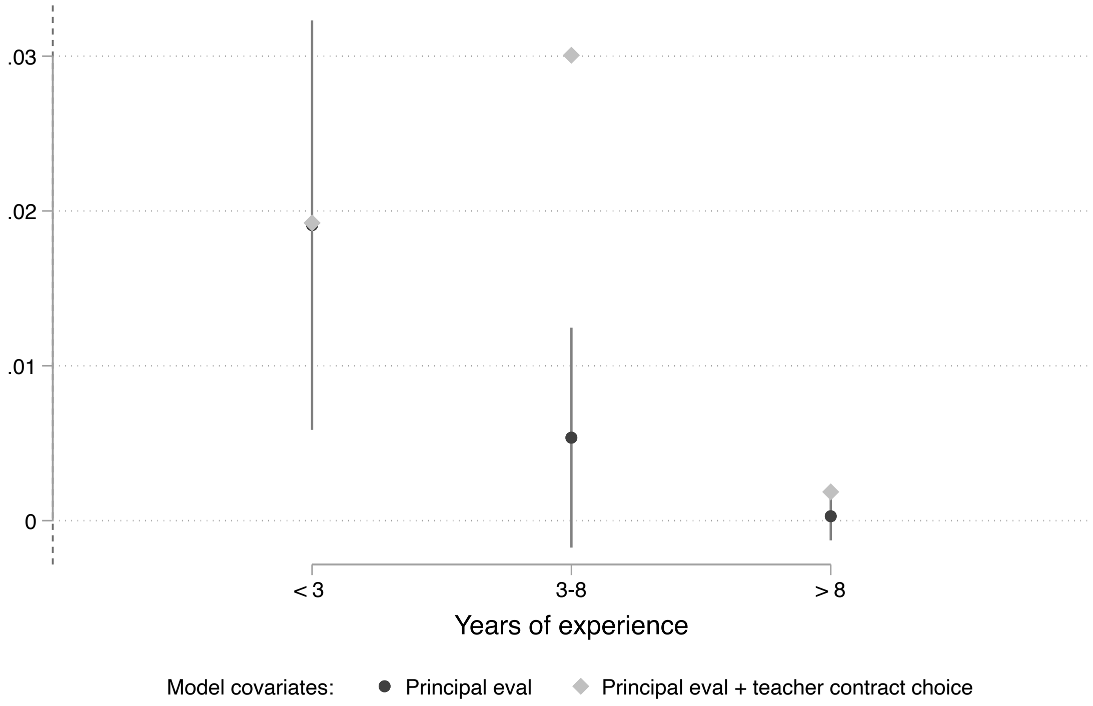
Notes: This figure presents the relationship between value-added and predicted value-added for three different models. The first model (solid line) just includes teacher demographics (age, experience and credential-type fixed effects). The second model (dashed line) uses demographics and principal evaluation.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

Figure A12: Treatment Distribution Map, Lahore



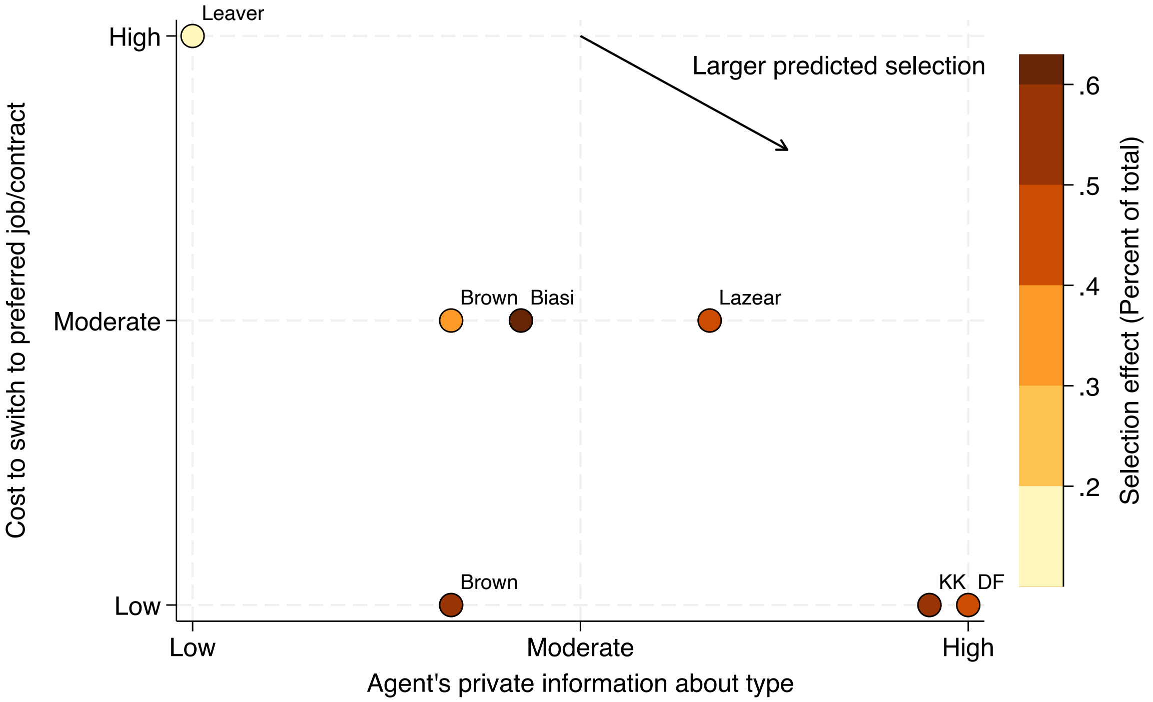
Notes: The figures shows the location of treatment versus control performance pay assignments in one of the cities in our study.

Figure A13: Predicting Teacher Value-Added by Experience



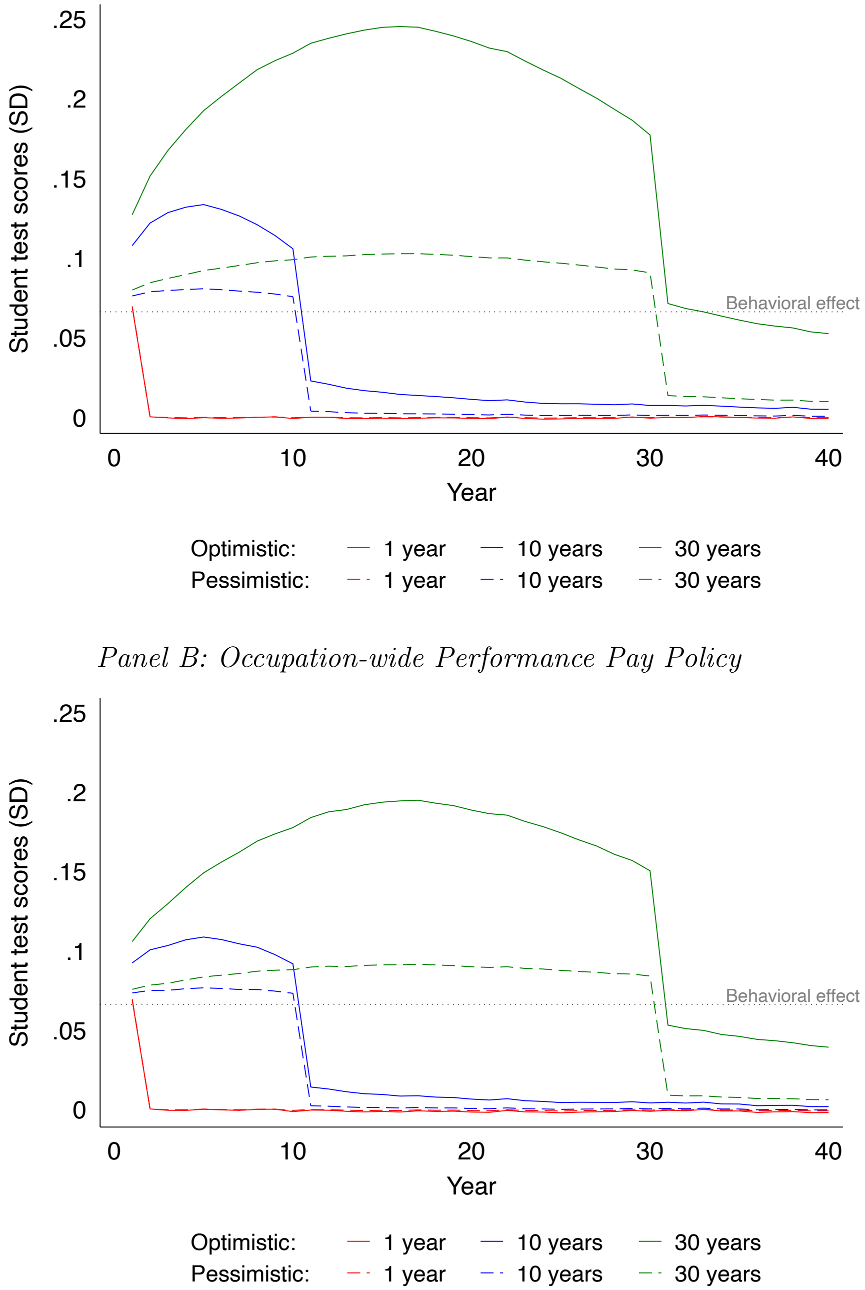
Notes: This figure presents the coefficient and 95% confidence intervals for predicted value-added on value-added for two different models. The first model (black circle) uses principal evaluation. The second (gray diamond) model includes principal evaluation and teacher's baseline contract choice. Results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

Figure A14: Comparison of Selection Effects Across Settings



Notes: This figure presents the extent of selection in response to performance pay from other studies. The x and y-axes show different aspects of the setting (extent of private information agents would have and how costly it would be to switch into their preferred contract). The color of the points correspond to what fraction of the total effect of performance pay was attributed to selection in the study. Based on our conceptual framework, we would expect larger selection effects as we move down and to the right. Table A10 describes each study in more detail.

Figure A15: Policy Simulations



Panel A: District-wide Performance Pay Policy

Panel B: Occupation-wide Performance Pay Policy

Notes: These figures presents the results of the policy counterfactual simulations on average test scores. Each panel 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. Panel A assumes the performance pay policy is introduced at a district level, and Panel B assumes the policy is a national change affecting the entire teaching occupation.

Table A1: 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, 2022 (PPP))	18,700	8,800	1,800	1,200	9,500	4,700
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 A2: Baseline Covariates

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

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 A3: Baseline Covariates - Information and Neighbor Treatment

Panel A: Information Treatment Balance

	(1)		(2)		T-test
	Control		Information Treatment		Difference
Variable	N/[Clusters]	Mean/SE	N/[Clusters]	Mean/SE	(1)-(2)
Performance evaluation score	1227	3.487	471	3.473	0.014
	[167]	(0.021)	[133]	(0.040)	
Salary (USD)	1468	5691.294	552	5682.607	8.687
	[180]	(123.891)	[146]	(125.564)	
Age	1469	38.608	553	38.908	-0.300
	[180]	(0.286)	[146]	(0.383)	
Years of experience	1468	6.834	552	6.973	-0.139
	[179]	(0.223)	[146]	(0.286)	

Panel B: Neighbor Treatment Balance

	(1)		(2)		T-test
	Same Treatment		Opposite Treatment		Difference
Variable	N/[Clusters]	Mean/SE	N/[Clusters]	Mean/SE	(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: These tables summarize teacher characteristics before the experiment by the two treatment dimensions used in section 6. Panel A compares teachers who received the information treatment versus no information. Panel B compares teachers whose neighboring school had the same performance pay treatment versus those whose neighboring school had the opposite treatment. The final column reports mean differences between treatment and control group. $^{*}p < 0.10, ^{**}p < 0.05, ^{***}p < 0.01$.

Table A4: 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. *Assigned Perf Pay Treat* is a dummy for whether a teacher taught at a school assigned to performance pay at baseline. *% Perf Pay* ranges from 0 to 1 and is the percent of their raise they wanted to be based on their performance. *Principal Rating of Teacher* is the principal's evaluation score of the teacher (as a standardized z-score). 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 A5: 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 measured using test score data from the three 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 A6: 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 A7: Positive Sorting by Closest School's Treatment

	Teacher Baseline VA (in Student SDs)	
	(1)	(2)
Performance Pay Schools	-0.0207	-0.0138
	(0.0330)	(0.0417)
Post	0.00243	-0.0267*
	(0.0228)	(0.0139)
Perf Pay Schools*Post	-0.00259	0.0373**
	(0.0232)	(0.0169)
Neighbor's Treatment	Same	Opposite
F-test p-value	0.149	
Control Mean	0.0190	0.0190
Control SD	0.327	0.327
Clusters	172	115
Observations	3495	1211
Randomization Strata FE	Yes	Yes

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 three 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 teachers whose closest neighboring school was assigned the opposite treatment as their school (low switching cost). Columns (2) presents the results for teachers whose closest neighboring school had the same treatment as them (high switching costs). *F-test p-value* tests for the equality of the *Perf Pay Schools*Post* coefficients across the two samples. Standard errors are clustered at the school level. $^{*}p < 0.10, ^{**}p < 0.05, ^{***}p < 0.01$.

Table A8: Treatment Effects on Classroom Observations by Contract Choice

	Classroom Observation Score				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 A9: Treatment Effects on Non-Test Student Outcomes 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. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table A10: Comparison of Selection Effects Across Other Studies

Study	Setting	Contracts	Outcome	Design	Total effect	Percent of total effect		Switching costs	Private info
						Selection	Behavioral		
Lazear (2000)	Manufacturing, Ohio	Piece rate vs flat	Units per day	Phased rollout	0.37	46%	54%	Moderate	High
Dohmen and Falk (2011)	Lab math task, Germany	Piece rate vs flat	Correct answers	Two stage RCT	31	43%	57%	Zero	Very high
Kantarevic and Kralj (n.d.)	Physicians, Ontario	Fee for service vs flat	Services per day	Dif-in-dif	1.03	58%	42%	Zero	Very high
Biasi (2021)	Teachers, Wisconsin	Flexible vs seniority pay	Teacher value-added	Dif-in-dif	0.09	63%	37%	Moderate	Moderate
Leaver et al. (2021)	Teachers, Rwanda	Performance vs flat pay	Test scores (SD)	Two stage RCT	0.02	20%	80%	High	Low
Brown & Andrabi (2024)	Teachers, Pakistan	Performance vs. flat raise	Test scores (SD)	Two stage RCT (contract choice)	0.14	53%	47%	Zero	Moderate
Brown & Andrabi (2024)	Teachers, Pakistan	Performance vs. flat raise	Test scores (SD)	Two stage RCT (job choice)	0.099	33%	67%	Moderate	Moderate

Notes: This table shows the findings from the prominent other studies which investigate the selection and effort effect of performance pay across any setting. Total effect and percent of total effect are calculated using authors results. Estimated switching costs and private information are assessed given the description of the setting and policy.

Table A11: Values of Key Parameters

Parameter	Value		Source/Calculation
	Pessimistic	Optimistic	
Mean ability, μ_θ	0	-	Baseline test scores
SD ability, σ_θ	0.15	0.30	Baseline test scores
Mean behavioral effect, μ_β	0.07	-	Experiment
SD behavioral effect, σ_β	0.14	0.28	Experiment
Covariance θ and β, $\rho_{\theta\beta}$	$-1.94\text{x}10^{-4}$	-	Experiment
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
Mean cost to change district, μ_c	\$560	-	Survey
SD cost to change district, σ_c	\$600	-	Survey
Accuracy of priors θ, α_θ	0.049	0.075	Experiment
Accuracy of priors β, α_β	0.025	0.038	Experiment
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.

B 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 district or state introducing performance pay and the effect of a more permanent policy. 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 for a short-term, medium-term and long-term performance pay policy, applied to *all* schools.

We augment the simple framework from section 2.1 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 each year by selecting the job which maximizes: i). the expected flow of future wages, w_{jt}, for their remaining time in the labor force, τ, ii). minus the cost to change sectors if the job is not in the sector the employee currently works in, and 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:

$$u_t = \underset{j}{max}\,\underbrace{(\sum_{t=1}^{T} \underbrace{[p(\alpha_\theta\theta_i + \alpha_\beta\beta_i)]\mathbb{1}[\delta_{jt} = 1]}_{\text{wage by contract type}}\,\mathbb{1}[\tau_i > t])}_{\text{over length of employment}}\underbrace{-c_i\mathbb{1}[s_t \neq s_{t-1}]}_{\text{cost to change sector}} + \underbrace{\epsilon_{ij} + \epsilon_{ijt}}_{\text{non-wage utility}}$$

Table A11 presents the key parameter values used. We use the covariance in α and β from the data, and assume the covariance between the other teacher characteristics is zero. To calculate the mean, variance and covariance 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 moments of θ and β and their covariance are calculated as follows:

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

Our estimates of σ^2_θ and σ^2_β come from the existing set of teachers in the school system. For our pessimistic values, we assume the variance in θ and β is the same in our sample as in the labor market in general. For the optimistic values, we assume the variance is twice as large in the full labor market given there are likely very high and low value-added individuals working outside the teaching profession. The fraction of new entrants each year (0.3), and the distribution of length of time on the job: $minexp(10), 50$ come from the administrative data.

The variance 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 teachers come directly from the contract choice experiment. We also include optimistic values of the parameters, which are 50% larger than in the current experiment 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 A15 presents the effects over time of introducing a 1 year, 10 year or 30 year performance pay policy at the district level (Panel A) or occupation level (Panel B).

The effect of a 1 year policy at the district-level is just the average behavioral response (0.07 SD) as there is virtually no sorting response. Under a 10 year policy, there is an average effect of 0.08 SD (0.12 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 0.10 SD (0.21 SD) under pessimistic (optimistic) parameters. These effects are 1.5-3.1x larger than the one year effects of performance pay. The effects also build over time, as better teachers sort in each year. We also find some persistence in effects once the policy is removed, as better teachers will have sorted in and switching costs are large enough to then keep them in the teaching sector even in the absence of performance pay.

The patterns for an occupation-wide policy follow a similar qualitative pattern but the magnitudes are slightly smaller, as there is less overall sorting. The effect of a 10 year policy ranges from 0.08 - 0.10 SD, and the effect of a 30 year policy is 0.09 - 0.17 SD.

C Tournament Structure

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. To properly model this contract, we would need to determine the appropriate equilibrium game play, as a single teacher's payoff now depends on other teacher's behavior. 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.

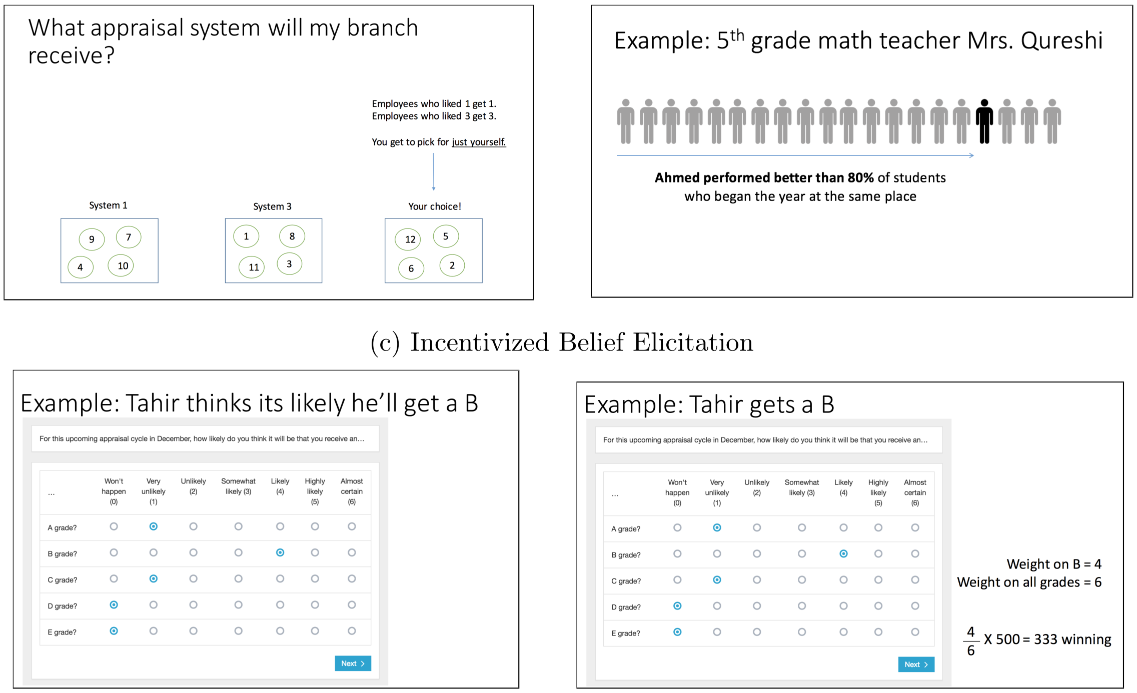
How does this affect our results? In the contract choice exercise, the tournament nature would not dissuade teachers from selecting performance pay because teachers are compared to all teachers at the school (not just the ones who selected performance pay). For the job choice results, our effects depend on how much switching teachers expected there to be. At endline, we surveyed teachers and asked them to predict 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 correctly predict the actual level of sorting we find (0.013 SD), this should only dissuade teachers between the 50th and 52nd percentile from sorting, which would result in sorting effects which are greater than 99% the size of effects if we ignored these discouragement effects.

D Experiment Details

D.1 Contract Choice Exercise and Baseline Survey

The baseline survey and contract choice exercise was carried out in a random sub-set of schools in October 2017. The research team visited each school and administered the 45-minute survey to a group of six teachers at a time. Teachers read and answered the survey question on their own, while the surveyor helped with any issues the teachers faced. The survey included videos to explain the key concepts (figure D.1).

Figure D.1: Screen capture from baseline survey



(a) Contract Randomization

What appraisal system will my branch receive?

Employees who liked 1 get 1.
Employees who liked 3 get 3.

You get to pick for just yourself.

(b) Value-Added Calculation

Example: 5th grade math teacher Mrs. Qureshi

Ahmed performed better than 80% of students who began the year at the same place

(c) Incentivized Belief Elicitation

Example: Tahir thinks its likely he'll get a B

For this upcoming appraisal cycle in December, how likely do you think it will be that you receive an...

...	Won't happen (0)	Very unlikely (1)	Unlikely (2)	Somewhat likely (3)	Likely (4)	Highly likely (5)	Almost certain (6)
A grade?	○	●	○	○	○	○	○
B grade?	○	○	○	○	●	○	○
C grade?	○	●	○	○	○	○	○
D grade?	●	○	○	○	○	○	○
E grade?	●	○	○	○	○	○	○

Example: Tahir gets a B

For this upcoming appraisal cycle in December, how likely do you think it will be that you receive an...

...	Won't happen (0)	Very unlikely (1)	Unlikely (2)	Somewhat likely (3)	Likely (4)	Highly likely (5)	Almost certain (6)
A grade?	○	●	○	○	○	○	○
B grade?	○	○	○	○	●	○	○
C grade?	○	●	○	○	○	○	○
D grade?	●	○	○	○	○	○	○
E grade?	●	○	○	○	○	○	○

Weight on B = 4
Weight on all grades = 6

$\frac{4}{6}$ X 500 = 333 winning

Notes: This figure shows a screen capture from the video explaining some of the key concepts. Panel A shows a screen capture from the video explaining to teachers how their contract choice would be implemented with some probability. Panel B shows a screen capture from the video explaining to teachers how percentile value-added was calculated, giving teachers practical examples. Panel C shows screen captures 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.

Table D.1: 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
2. If I really try hard, I can get through to even the most difficult or unmotivated students	Efficacy	RAND Teacher Efficacy
3. "Smartness" is not something you have, rather it is something you get through hard work	Efficacy	RAND Teacher Efficacy
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
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
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	Ashraf et. al. (2020)
13. It is important to me to do good for others through my work	Intrinsic Motivation	Ashraf et. al. (2020)
14. I want to help others through my work	Intrinsic Motivation	Ashraf et. al. (2020)
15. One of my objectives at work is to make a positive difference in other people's lives	Intrinsic Motivation	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	Ashraf et. al. (2020)
17. My students matter a great deal to me	Intrinsic Motivation	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".

The survey was conducted in partnership with the school system's HR and administrative staff to ensure high perceived legitimacy of the process. Teachers were assured their responses to all questions would remain anonymous. We communicated that the one exception to this would be in the case their school was selected to implement the contract the teacher chose from themselves. In this instance, then their manager would need to be told what contract they chose in order to implement that choice.

In addition to the contract choice questions, teachers were asked questions assess their self-efficacy, intrinsic motivation, pro-sociality and long term career plans. Table D.1 presents the question items and item source.

D.2 Contract Randomization

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. Information about the contract they had been assigned was available on their employment dashboard. 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.

D.3 Classroom Observation

Prior to the rollout of classroom observations we conducted a survey and pilot to test different observation modalities (in person, video recording, video + audio recording). Video + audio recording was chosen based on being both the most reliable, cheapest and most preferred by teachers. Video-taping was perceived by teachers as less intrusive than human observation (and hence preferred by teachers). Video-taping also allowed for ongoing measurement of inter-rater reliability (IRR) with the team reviewing the videos at our research office.

Video and audio recording was done by setting up a tablet on a tripod in the back of the classroom for a five hour period. Audio recording devices were placed on a table at the front the classroom near where the teacher said they typically stood when speaking to the class.

The videos were reviewed and coded using the CLASS rubric by a team of local enumerators. We did not hire the Teachstone staff (the organization that produces the CLASS rubric) 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 inter-rater reliability (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.

D.4 Endline Test and Survey

At endline, students completed assessments in core academic subjects and completed a survey to assess five areas of socio-emotional development (love of learning, ethics, global citizenship, inquisitiveness and whether they like their school). The student survey items and source are shown in table D.2. Students were told their responses in both the assessments and survey would be anonymous.

Exam papers were delivered to schools in sealed envelopes which were to remain sealed until the time of the exam. Student's home room teacher proctored the exam, which would be a different teacher than the subject they were being tested in. Research staff collected the exams at the end of the testing week and conducted the grading of the exams. Research staff also did random drop in spot checks in classrooms during the exams themselves. There were no reports from these spot checks of teachers trying to help students in anyway (giving answers, showing test papers ahead of time, changing answers afterwards, etc).

Table D.2: 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 (children's)
5. I feel sorry for other kids who don't have toys and clothes	Ethics	Eisenberg's Child-Report Sympathy Scale
6. Seeing a child who is crying makes me feel like crying	Ethics	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)	Ethics	Bryant's Index of Empathy Measurement
8. The pressure to do well is very high, so it is ok to cheat sometimes (reversed)	Ethics	Bryant's Index of Empathy Measurement
9. I am interested in public affairs	Global citizenship	Afrobarometer/World Values
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 citizenship	Afrobarometer
11. People who are poor should work harder and not be given charity (reversed)	Global citizenship	Afrobarometer
12. It is important to protect the environment even if this means we cannot consume as much today	Global citizenship	Afrobarometer
13. People from other places can't really be trusted (reversed)	Global citizenship	Afrobarometer
14. I am comfortable asking my math/science/Urdu/English teacher for help or support	Inquisitiveness	Learning and Study Strategies Inventory
15. I enjoy learning about subjects that are unfamiliar to me.	Inquisitiveness	Litman and Spielberger, Epistemic Curiosity
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.