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

Christina Brown† Tahir Andrabi

October 30, 2020

Latest version of the paper here

Abstract

Attracting and retaining high-quality teachers has a large social benefit, but it is challenging for schools to identify good teachers ex-ante. This paper uses teachers' contract choices and a randomized controlled trial of performance pay with 7,000 teachers in 243 private schools in Pakistan to study whether performance pay can attract and retain higher-quality teachers. Consistent with adverse selection models, we find that performance pay can induce positive sorting: both high value-added teachers and teachers who respond more strongly to incentives significantly prefer performance pay and sort into these schools. Using two additional treatments, we show effects are more pronounced among teachers with more information about their quality and teachers with lower switching costs. Teachers have more information about their quality than their principals, and this holds throughout most of their tenure. Accounting for these sorting effects, the total effect of performance pay on test scores is twice as large as the direct effect on the existing stock of teachers, suggesting that analyses that ignore sorting effects may substantially understate the benefits of performance pay.

*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, Gautam Rao, Jesse Rothstein, Heather Schofield, and seminar audiences at UC Berkeley, Pacdev, CERP, RISE, and Stanford provided helpful feedback. We gratefully acknowledge generous funding and support by DFID's RISE Programme, JPAL's Post-Primary Initiative, 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, Zahra Niazi, Mujahid Murtaza, Maheen Rashid, and Zohaib Hassan, 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 California, Berkeley (christinabrown@berkeley.edu); Andrabi: Lahore University of Management Sciences and Pomona College

1 Introduction

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

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

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

Our experiment proceeds in two phases. First, we offer teachers the opportunity to choose their preferred contract, selecting among pairwise comparisons of a flat raise versus a performance-based raise. Teachers' choices are implemented in a randomly selected subset of schools to ensure incentive compatibility of responses. We also elicit the distribution of teachers' beliefs about their value-added and risk preferences through an incentivized activity.

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

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

Overall, we find strong evidence that performance pay induces positive sorting among high performing teachers. First, we find that teachers who choose performance pay contracts have higher value-added. Contract choice is predictive of value-added even when controlling for principal's information about teachers. These results are strongest among teachers in the middle of their careers (6-10 years of experience).

Second, we find positive sorting along actual job choice. The composition of teachers in performance pay schools is better after one year. These effects are mostly driven by high value-added teachers moving from control to treatment schools and low value-added teachers moving from treatment to control schools. High value-added teachers are also slightly more likely to leave control schools to work outside this network of schools. We do not find any effect on new entrants to the school system.

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

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

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

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

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

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

Second, we add to a robust literature on the direct, motivational effect of performance pay for teachers by providing two new findings (Lavy, 2007; Muralidharan and Sundararaman, 2011; Fryer, 2013; Goodman and Turner, 2013). We show that there is substantial heterogeneity in the direct effect of performance pay across teachers. Specifically, teachers who want performance pay have much larger behavioral responses than those that do not want performance pay. This suggests that in the long run, the effects of performance incentives could be much larger than the short term effects previously estimated. In addition, this behavioral response appears to be unrelated to baseline value-added. This suggests that the marginal effort response to incentives is uncorrelated with the equilibrium effort under no incentives.

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

Leaver et al., 2019; Biasi, 2017).

The remaining sections are organized as follows: Section 2 provides context about the use of performance pay in teaching. Section 3 presents the motivating model in the vein of Roy (1951). Section 4 details the contract choice elicitation, randomized controlled trial, and data collection procedures. Section 5 presents the results on the extent of positive sorting in response to performance pay, and Section 6 describes the extent of information principals have about teachers. Section 7 presents results on the sensitivity of the magnitude of positive sorting to teacher's switching costs and information, and Section 8 examines whether there is sorting along negative characteristics. Section 9 presents results from a policy simulation exercise.

2 Teacher Quality, Labor Market and Performance Pay

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

Despite the importance of teacher quality, schools have limited capacity to screen in and retain good teachers and screen out and lay-off bad teachers, due to institutional and information constraints. Public schools are typically severely constrained in their ability to fire bad teachers. Furthermore, it is not clear that schools can even identify who the high and low performing teachers are, either at the time of hiring or throughout the teacher's tenure. Characteristics available to schools at the time of hiring, including interview scores, explain less than 5% of teacher value-added (Bau and Das, 2020; Staiger and Rockoff, 2010; Rockoff and Speroni, 2010). Schools could potentially exploit teachers' private information about their quality by offering performance pay and causing high-quality teachers to self-select in. Lazear (2000) shows that employees in a glass factory positively sort in response to performance pay, and sorting effects are twice as large as the effects on effort.

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

Understanding the full effects of performance pay including both direct effects on existing teachers and sorting effects is crucial, as there has been a significant push to tie teacher salaries to student outcomes in developed and developing countries (Goodman and Turner, 2013; Pham et al., 2020; Muralidharan and Sundararaman, 2011). Across the world, the number of countries that use performance incentives for teachers doubled in the last decade, from one-third to two-thirds (World Bank, 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.

3 A Model of Job Choice

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

3.1 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^2_θ and σ^2_β, respectively, and covariance $\rho_{\theta,\beta}$.

The wage from each contract is then:

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

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 $\hat{\theta}_i = \alpha^\theta_i \theta_i + (1 - \alpha^\theta_i)\mu_\theta$ and $\hat{\beta}_i = \alpha^\beta_i \beta_i + (1 - \alpha^\beta_i)\mu_\beta$, where $\alpha^{\theta,\beta}_i \in [0, 1]$. The parameters α^θ_i and α^β_i govern teachers' prior certainty about ability and treatment effects. An α of 1 is perfect information about their ability or behavioral effect. An α of 0 implies that teachers have no information about their own ability or behavioral effect, and so their prior shrinks to the population mean.

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 may also gain non-wage utility from the type of contract they receive, such as disliking inequality or enjoying competition. However, in section 8.2, 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:

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

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

$$b_i = v_{iP} + \epsilon_{iP} - (v_{iF} + \epsilon_{iF}) \tag{3}$$

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

3.2 Employee Quality by Job Type

We treat employment as a one-sided job choice by the employee. Employers accept anyone that applies to the firm.[1] 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, β, for the subset of teachers who chose performance pay ($b \geq 0$).

The difference in average output for firms that offer performance pay versus flat pay then is:[2]

$$\Delta\bar{y} = E[\theta_i + \beta_i | b_i \geq 0] - E[\theta_i | b_i < 0] \tag{5}$$

$$= \underbrace{E[\theta_i | b_i \geq 0] - E[\theta_i | b_i < 0]}_{\text{sorting on ability}} + \underbrace{(E[\beta_i | b_i \geq 0] - E[\beta_i | b_i < 0]) P(b_i < 0)}_{\text{sorting on behavioral effects}} + \underbrace{E[\beta_i]}_{\text{avg. behavioral effect}} \tag{6}$$

The first term, "sorting on ability", captures the difference in average underlying ability between those who choose performance pay versus those who do not. The second term, "sorting on behavioral

[1] Section 5.1 will show this is a reasonable assumption in our setting. We will also relax this constraint by presenting results controlling for principal information to mimic settings where principals can screen employees.

[2] Proof in Appendix B.

effect" represents the difference in behavioral response to incentives for those who choose performance pay versus flat pay. Together these two terms comprise the sorting effect of performance pay contracts, which together we will refer to as Δy_s. The last term ("average behavioral effect") captures the average behavioral response to performance pay for all teachers. This term is the effect of performance pay contracts on the static population of teachers, similar to what other studies of performance pay have focused on. Our focus for this paper will be to estimate both the sorting effects (the first two terms) and the direct behavioral effects (last term).

3.3 Model Predictions

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

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

To test each of these predictions, we conduct a randomized controlled trial. A key assumption of the model is that non-wage utility from a job is independent of the contract. In our experiment, that assumption is satisfied by randomizing performance versus flat pay contracts across schools, allowing us to test predictions 1. In addition, we exogenously vary teachers' information about their ability via an information treatment and the variance of non-wage utility by varying the distance between jobs with opposite contract treatments, allowing us to test predictions 2 and 3.

4 Experimental Design

4.1 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 October 2017 to June 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 with 2,480 teachers. Teachers were asked to choose between several contracts for the following year and told that the contract they chose would be implemented with some probability. The implied likelihood from the survey was that there would be a one-third chance their choice would be implemented.[3] Teachers were asked about two sets of choices: i). flat raise contract versus performance raise contract based on an objective measure of performance (percentile value-added), ii), flat raise versus performance raise based on a subjective measure of performance (principal evaluation).

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

Phase 2: Contract Randomization To measure the behavioral effects of performance pay, we randomize contracts across the remaining 243 schools that were not selected to implement the teacher's contract choice. Schools were randomized to receive one of three contracts that determine the size of teachers' raises at the end of the calendar year.[4,5] The three contracts were:

- **Control: Flat Raise** - Teachers receive a flat raise of 5% of their base salary.
- **Treatment: Performance Raise** - Teachers receive a raise from 0-10% based on their within-school performance ranking.[6]

[3]Appendix figure C3 presents information about how this probability was explained to participants, including screen captures from the video shown to participants. The actual implementation probability was a bit lower than one-third due to implementation constraints.

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

[5]To ensure teachers fully understood their contract, we conducted an intensive information campaign with schools. First, the research team had an in-person meeting with each principal, explaining the contract assigned to their school. Second, the school system's HR department conducted in-person presentations once a term at each school to explain the contract. Third, teachers received frequent email contact from school system staff, reminding them about the contract, and half-way through the year, teachers were provided midterm information about their rank based on the first six months. An example midterm information note is provided in appendix figure C5. 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.

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

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

There are two treatment sub-arms, which vary the performance measure used to evaluate teachers. Teachers are ranked within their school on either:[7]

– ***Objective Performance:*** Percentile value-added (Barlevy and Neal, 2012) averaged across all students they taught during the spring and fall term.[8]

– ***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.[9]

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

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

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

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

[8]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.

[9]These included items such as improving their behavioral management of students, assisting with administrative tasks, helping plan an after-school event, and improving students' spoken English proficiency. An example set of criteria are provided in appendix figure C4.

4.2 Data

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

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

Teacher and principal survey In addition to the contract choice exercise, the baseline survey included incentivized measures of teacher's beliefs about their performance along the objective (percentile value-added) and subjective (principal evaluation) metric. We also measured teachers' risk preferences using a high-stakes (a week's wage) and medium-stakes (half a day's wage) coin flip game and pro-sociality using responses to a volunteer opportunity. 40% of schools were randomly selected to participate in the baseline survey (and contract choice exercise). Data collection was conducted in October 2017, three months before the announcements of treatments.

At endline, we again measure teacher beliefs about their value-added, risk preferences, and offer a medium-stakes contract choice exercise. The survey also included measures of intrinsic motivation (Ashraf et al., 2020), efficacy (Burrell, 1994), and checks on what teachers understood about their assigned contract. The endline survey was conducted online with teachers and managers in spring and summer 2019. Appendix table C2 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.[10]

Endline Student Testing and Survey: An endline test was conducted in January to measure performance in Reading (English and Urdu), Math, Science, and Economics in grades 4-13.[11] The items were written in partnership with the school system's curriculum and testing department to ensure the appropriateness of question items. The research team conducted the grading. Items from international standardized tests (TIMSS and PERL) and a locally used standardized test (LEAPS) were also included to benchmark student performance. Students also completed a survey to measure four areas of socio-emotional development chosen based on the school system's student development

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

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

priorities.[12]

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

4.3 Measuring Teacher Ability

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

$$y_{ijkcst} = \beta_0 + \sum_s \beta_s y_{ijkcs,t\text{-}1} \mathbb{1}[\textit{subject-grade} = s] + \sum_s \alpha_s y_{ijkcs,t\text{-}2} \mathbb{1}[\textit{subject-grade} = s] \tag{7}$$
$$+ \sum_s \gamma_s \bar{y}_{\text{-}ijkcs,t\text{-}1} \mathbb{1}[\textit{subject-grade} = s] + \chi_{st} + \psi_k + v_{ijkcst}$$
$$\textit{where } v_{ijkcst} = \mu_j + \theta_{ct} + \epsilon_{ijkcst} \tag{8}$$

where y_{ijcst} 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

[12]The areas are (i). love of learning (items drawn from National Student Survey, Learning and Study Strategies Inventory), (ii). ethical (items from Eisenberg's Child-Report Sympathy Scale, Bryant's Index of Empathy Measurement), (iii.) global citizen (items from Afrobarometer; World Values Survey), and (iv.) inquisitive (items from Learning and Study Strategies Inventory; Epistemic Curiosity Questionnaire). Appendix table C1 lists the survey items used for each area along with their source. These are the four socio-emotional development areas they expect their teachers to focus on. These areas are posted on the walls in schools, and teachers receive professional development in these areas. Some principals also specifically make these areas part of teachers' evaluation criteria. In addition to four areas, the survey asked whether students liked their school.

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

[14]We did not hire the Teachstone staff to conduct official CLASS observations as it was cost-prohibitive, and we required video reviewers to have Urdu fluency. Instead, we used the CLASS training manual and videos to conduct an intensive training with a set of local post-graduate enumerators. The training was conducted over three weeks by Christina Brown and a member of the CERP staff. Before enumerators could begin reviewing data, they were required to achieve an IRR of 0.7 with the practice data. 10% of videos were also double reviewed to ensure a high level of 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.

coefficients on lagged test scores (β_s, α_s and γ_s) to vary across subject-grade. χ_{st} captures subject-grade-year shocks. ψ_k captures school-specific shocks. The residual, v_{ijkcst}, is the combination of teacher effects μ_j, classroom effects, θ_{ct}, and student-time specific shocks, ϵ_{ijkcst}. To isolate the teacher component, we use the residuals, v_{ijkcst}, to construct an empirical Bayes estimate of teacher value-added. We compute the average weighted residual and shrink by the signal variance to total variance ratio (Kane and Staiger, 2008).[15] Teachers for which we have few student observations are shrunk toward the mean teacher value-added (normalized to be zero).[16]

Having a teacher with a 1 SD higher VA for one year is associated with a 0.15 SD higher student test score. The effects are slightly larger for math, English, and Urdu and smaller for science. These effects are similar to other estimates from South Asia (0.19 SD, Azam and Kingdon (2014) and 0.15 SD, Bau and Das (2020)). Figure 2 shows the distribution of teacher value-added for the 3,687 teachers who teach in the school system at baseline.

4.4 Sample and Intervention Fidelity

Teacher and Principal Sample The study was conducted with a large, high fee private school system in Pakistan. The student body is from an upper middle-class and upper-class background. School fees are \$2,300-\$4,300 USD (PPP). Table 1, panel A, presents summary statistics for our sample teachers compared to a representative sample of teachers in Punjab, Pakistan (Bau and Das, 2020). Our sample is mostly female (81%), young (35 years on average), and the median experience level is 10 years, but a quarter of teachers are in their first year teaching. Nearly all teachers have a BA, and 68% have some post-BA credential or degree. Teachers are generally younger and less experienced than their counterparts in public schools, though they have more education. Salaries are, on average, \$13,000 USD (PPP). Yearly turnover is 29%. There is a mix of career teachers and those who are less attached to their school. 70% and 36% expect to still be teaching at their current school in 1 year and 10 years, respectively. Panel B presents information about sample schools and principals compared to a representative sample of schools in India (data was unavailable for Pakistan) (Bloom et al., 2015). Principals in our sample are more likely to be female and have much higher personnel management, operations, and performance monitoring scores than the average school in India.

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

[16] 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 two years of data. Schools are also relatively large, with an average of 131 students per grade. Students are tested every year, beginning in 4th grade.

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

Schools in the two treatment arms and control appear to be balanced along baseline covariates. Appendix table A5 compares schools along numerous student and teacher baseline characteristics. Of 27 tests, one is statistically significant at the 10% level, and one is statistically significant at the 5% level, no more than we would expect by random chance. Results control for these few unbalanced variables.

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

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

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

5 Positive Sorting

We now present the main results of the paper in sections 5 through 8. In this section, we present evidence on Prediction 1. We first show that higher value-added teachers are more likely to choose performance pay contracts compared to flat pay when they are allowed to select their contract for the following year. We then show higher value-added teachers are more likely to move into performance pay schools after contracts have been randomized across schools. Finally, we document larger direct treatment effects for teachers who chose performance pay.

5.1 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?".*[17]

We ask this question twice: once for an objective performance metric (percentile value-added) and once for a subjective performance metric (principal evaluation). Appendix C provides the full question description, including the examples given, understanding checks preceding the question, and explanation to teachers about how percentile value-added is calculated.

Figure A4 shows the distribution of teachers' responses. Most teachers want at least part of their raise to be performance-based, with less than 10 choosing a completely flat raise. On average, teachers wanted 56% of their raise to be performance-based when the performance metric was subjective and a slightly lower 52% when the performance metric was objective. For ease of communication going forward, we will group responses that are greater than 50% flat as "chose flat pay" and less than or equal to 50% as "chose performance pay". As an alternative, the appendix presents results treating the choice as a continuous variable. All of the main results are unchanged between the two approaches.

Figure 3 presents the relationship between contract choice and teacher demographics, characteristics, and beliefs. A strong predictor of contract choice is the teacher's belief of their principal's rating of them in the next year. Teachers that are more risk-loving (as measured in a real-stakes coin flip game) and those that say they are likely to stay teachers over the next five years also prefer performance pay. Female teachers are less likely to choose performance pay, and experienced teachers are slightly more likely to choose performance pay. These relationships generally hold whether the performance metric is subjective or objective (shown in Figure A2).

Positive Sorting in Contract Choice We find that teachers who chose a performance pay contract have significantly higher baseline value-added. Figure 4 plots the distribution of baseline value-added (in student standard deviations) for teachers who chose performance pay (solid line) versus those who chose flat pay (dashed line). The entire distribution is shifted to the right for those who wanted performance pay, and the difference is equivalent to a 0.03 SD difference in test

[17] 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.

scores. This difference holds for the choice between objective performance pay versus flat pay and subjective performance pay versus flat pay.

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

$$VA_{i,t-1} = \beta_0 + \beta_1 ChosePerfPay_i + \epsilon_i \tag{9}$$

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

Table 2 presents the results from eq. 9. As we showed in the figures, teachers who chose performance pay had 0.03 standard deviation higher baseline value-added. The relationship is similar whether we look at choices on objective or subjective performance pay. Columns (2) and (4) control for the principal's evaluation of the teacher. We see that principals do have some information about teacher value-added. A 1 SD increase in principal rating is related to a 0.014 SD increase in value-added. However, when we control for the information that principals have, the teacher's choice of performance pay is still a significant predictor of value-added. This suggests that teachers have additional information about their own quality beyond what principals know.

While on average teachers seem to have information about their ability, we do see heterogeneity across teacher type. Figure 5 presents the relationship between baseline value-added and likelihood of choosing performance pay by teacher gender, age, and experience. Here a steeper line suggests more positive sorting in response to performance pay. The average level of the line shows the extent to which performance pay is preferred on average for that sub-group. First, we see female teachers are less likely in general to prefer performance pay but have a similar relationship between ability and contract choice as male teachers. We also see that more novice teachers appear to have less information about their ability or, at least, are not sorting on that information. However, we also see that older teachers may be more overconfident and their abilities and, therefore, more likely to choose performance pay even when they are not actually high ability.

Measuring Job Choice Next, we investigate whether the composition of teachers changes between flat pay versus performance pay schools. We use administrative data from the school system to identify where each individual works at baseline (December 2017) and a year after the contracts are announced (December 2018). We observe if a teacher joins or leaves the school system but do not know if and where they are employed if they leave the school system.[18,19] During the treatment information campaign, teachers were also told if they transferred schools, they would be subject to

[18] 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.

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

the contract of the school they transferred to.[20] Transfers are initiated by the teacher and need to be accepted by the receiving school.[21] Transfers are nearly always accepted by the receiving school. This is because incumbent teachers have hiring priority, and there is high turnover within the system, virtually guaranteeing open positions at the school of interest each summer. Therefore it is appropriate to think of this setting as a one-sided choice problem, as the schools have little say in who within the transfer applicants is hired.

Positive Sorting in Job Choice Figure 6 presents the distribution of teacher value-added at baseline (Panel A) and then one year after the announcement of the contract (Panel B) across treatment and control schools. At baseline, the two distributions are virtually indistinguishable. However, a year later, there are now more below-average value-added teachers in flat pay schools and more above-average value-added teachers in performance pay schools, with an average difference of 0.016 SD. Similarly, we can see the cumulative distribution functions lie on top of each other at baseline, but, a year later, the performance pay schools dominate flat pay schools at every part of the distribution (figure A8).

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

$$VA_{i,t-1} = \beta_0 + \beta_1 WorkatPP_i + \beta_2 Post_i + \beta_3 WorkatPP_i * Post_i + \chi_j + \epsilon_i \tag{10}$$

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

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

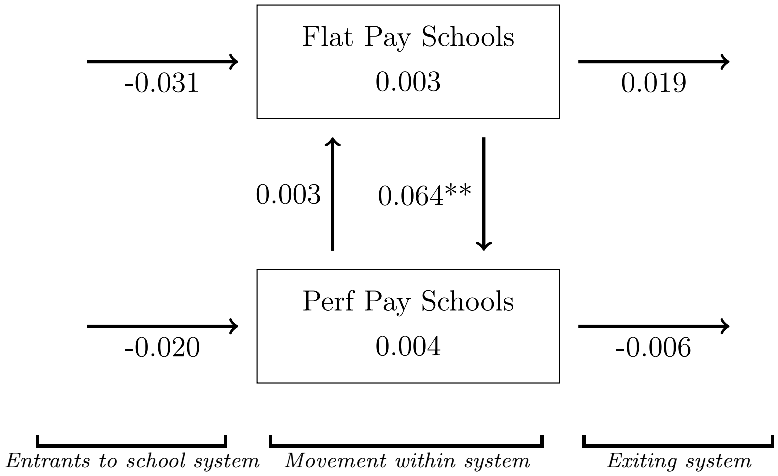
[21] There are two types of transfers. Many schools operate on a larger campus. For example, there may be a primary school, middle school, and high school all on the same larger campus, and a teacher applies to transfer from the primary school to the middle school. For example, the other type is across campuses transferring from a middle school teacher at a school in Lahore to a different branch of the school system in Karachi. 6% of teachers make a within campus transfer, and 11% of teachers make an across campus transfer each year. Transfers are recorded in the administrative data, and we can observe rejected transfer applications. The vast majority of transfers and resignations happen over the summer break between school years (calendar of transfers shown in figure A3).

Table 4, column 1, presents the results of eq. 10. As we saw with the figures, there is no difference between performance and flat pay schools at baseline. However, a year later, the average baseline value-added of teachers at flat pay schools is 0.013 SD lower in flat pay schools and 0.002 SD higher in performance pay schools (a difference of 0.015 SD between treated and control schools). The magnitude of this effect is relatively small, but as this was just a one-year contract change, it is not surprising we do not find huge shifts in employment across schools. As this is the extent of positive sorting from a one-year contract change, we would expect this to be a lower-bound on the extent of sorting.

The results are robust to additional controls in columns 2 and 3 for region, grade, and subject. Column 4 adds controls for the principal's rating of the teacher. Principals appear to have some information about teacher quality. A 1 SD increase in the principal's rating of the teacher is associated with a 0.06 SD higher teacher value-added (0.0134 SD in student standard deviations). However, the coefficient on $WorkatPP_i * Post_i$ remains significant when we control for principal information, so this sorting behavior is providing a signal about teacher's quality beyond what principals know already, suggesting teachers do have private information. We do not see any significant differences in sorting by gender, age, or experience.

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

The diagram below maps the change in teacher quality for teachers who switch within the system, leave the system, and are new entrants to the system during the intervention year. The numbers next to each arrow show the average baseline value-added for that group. For example, the arrow in the top left part of the diagram shows that the average value-added for teachers who are entering the school system and starting their first job at a flat pay school is -0.046 SD. The numbers inside the boxes show the average value-added for teachers who stayed at their original school or moved from a school to another school with the same treatment. For example, teachers who stayed at a flat pay school or moved from one flat pay school to another flay pay schools had an average baseline value-added of 0.003 SD.



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

5.2 Positive Sorting on Behavioral Effect

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

$$\begin{aligned} TestScores_i = \beta_0 + \beta_1 AssignedPPtreat_j + \beta_2 ChosePP_i \\ + \beta_3 AssignedPPtreat_j \cdot ChosePP_i + \beta_4 TestScore_{i,t-1} + \chi_j + \epsilon_i \end{aligned} \tag{11}$$

The outcome is endline test scores for students taught by teacher, i. $PPtreat_j$ captures the treatment assigned to the teacher's school, j for the school at which the teacher taught at the time of treatment announcement. As we saw in section 5.1, some teachers change schools during the experiment, so $PPtreat_j$ gives us the intent-to-treat effects of performance pay. $ChosePP_i$ is the teacher's contract choice from the baseline survey. We control for randomization strata, χ_j, and student's baseline test scores, $TestScore_{i,t-1}$. Standard errors are clustered at the school level (the unit of randomization). The coefficient of interest is β_3, which captures whether there is a differential effect of performance pay on teachers who wanted that contract. We, of course, restrict to the RCT sample of schools, so the $ChosePP_i$ variable is unrelated to the contract assigned, $AssignedPPtreat_j$.

We find that teachers who wanted performance pay have much larger behavioral responses than those who wanted flat pay, (0.09 SD versus 0.01 SD). Figure 7 presents the average effect of performance pay across all teachers and then splits the sample by teachers who chose performance pay versus those who chose flat pay. Table 5, column 2, presents the results of equation 11. Column 3 controls for principal rating, which does not change our effects. In fact, along this metric we do not find that principals have information about teacher quality. Results, shown in table A2, are also identical if we treat contract choice as a continuous variable (percent of raise chose to be performance-based).

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

Total Effect of Performance Pay Returning to our decomposition of the total effect of performance pay eq. 5, we have the following total effect:

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

We summarize the effect of each of these components in the setting without switching costs (contract choice exercise) and with high switching costs (teacher job choice in the second year). When we incorporate sorting effects, we see that the total effect of performance pay is somewhere between 86% and 44% larger than measuring just the effect on the existing stock of teachers.

6 Asymmetric Information

6.1 How much information do employers have?

As we saw in table 2 and 4, principals do have some information about teacher quality. However, the extent of principal information varies substantially depending on the dimension of teacher quality and principal's exposure to teachers. At endline, we ask principals to rate teachers they oversee along four dimensions of quality: i). attendance, ii). managing student discipline in the classroom, iii). incorporating higher-order skills in lessons, such as analysis and inquiry, and iv). value-added. We then compare this to teachers' actual daily attendance, recorded via biometric clock in/out data, teachers' management of student discipline, and incorporation of higher-order skills assessed using classroom observation data, and teachers' actual value-added.

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

We also find that principal accuracy varies substantially depending on the level and type of exposure principals have with teachers. From September 2018 to January 2019, we randomly assign some teachers to receive more frequent classroom observations from their principals. Principals were instructed to observe treated teachers at least once a month during the period, though not all principals completed the full set of observations. We find that treated teachers receive 2.7 observations during the 5-month period, relative to 1.8 for the control.

Principals provide much more accurate ratings for teachers who were assigned to the observation treatment. Table 6, column 7, provides principal rating by observation treatment status. A 1 SD increase in teacher outcomes is associated with a 0.06 SD increase in principal rating for control teachers versus 0.25 SD for treated teachers. This increase in accuracy comes both from increasing their rating of high performers and lowering their rating of low performers.

However, principals actually get *less* accurate the longer they work with a teacher. Table 6, column 8, compares principal accuracy for principals who have worked at the same school as the teacher for more than or less than two years.[22] 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.[23] 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 effect is actually about overlap or something correlated with it. For example, the amount of time overlapping would also correlate with principal experience and job change frequency. While we cannot address every possible omitted variable, column 9, controls for principal and teacher years of experience, and column 10 controls for principal fixed effects. Our results are robust to the addition of these controls.

[22] 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.

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

6.2 How much more information do teachers have?

The policy-relevant parameter is how much more information teachers have than their employers. 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 8 plots predicted teacher value-added relative to actual value-added for each of these models. The solid line is from predicted value-added using age, experience, and credential-type fixed effects. We see that these criteria predict some variation in teacher value-added. The dashed line adds principal evaluation data to the model, which slightly improves the model (though we cannot reject equality of the two models). Finally, adding in teacher contract choice (dotted line) triples the predictive power of the model. This suggests that teachers have substantially more information about their type than their employer.

We find the extent of asymmetric information varies over a teacher's tenure. Figure 9 presents the coefficient on the regression of predicted value-added on actual value-added. The solid black circles and 95% confidence intervals show the coefficient when predicted value-added is constructed using just principal evaluation data. The gray diamonds show the coefficient when we add teacher contract choice to the prediction. The data is split by novice (less than 3 years), experienced (3-8 years), and very experienced teachers (greater than 8 years). We see an interesting pattern across teacher experience. As we showed in the effect of overlap with a teacher, principals become less accurate the more experienced a teacher is. Teachers initially become more accurate with experience but drop off for very experienced teachers. Teachers have more information than principals in all years except for very novice teachers.

What is the source of teacher's private information? There are two possible explanations for this result: (i) teachers have information about their own ability or (ii). teachers do not have information about their value-added, but value-added is correlated with other preferences (risk, competitiveness, etc.) that make high types more likely to choose performance pay. We do not find evidence for the second claim. Higher value-added teachers and those that have larger behavioral responses do not have different risk preferences, preferences for competition, or pro-sociality (table A3). We can also control for risk preferences, preferences for competition, and pro-sociality in our main positive sorting results on ability and behavioral effect (table A4). Our results remain unchanged when we control for these potential channels.

7 Magnitude of Positive Sorting

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

7.1 Sorting by teacher information

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

First, for this information treatment to work, teachers must not be fully informed about their own value-added. We find that teachers update in response to this information treatment. Figure 11, panel A, plots teacher's predictions about their performance in the coming year relative to their true performance that year for teachers who received no information versus those who learned about their historical value-added. Those that receive information do a better job of being able to predict their future value-added. This information also influences their ultimate contract choice. The correlation between choosing performance pay and teachers increases by 50% for those assigned to the information treatment versus no information, as we see in figure 11, panel B. This suggests that better information about one's own ability does increase the extent of positive sorting.

7.2 Sorting by switching costs

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

Most schools operate on a larger campus, which contains multiple schools (primary school, middle school, high schools). Within the same campus, different schools may be assigned to different contracts. Therefore, we can look at the extent of positive sorting when another school on the same campus was assigned to the opposite treatment as the teacher's own school's treatment. For example, we can see that in one of the cities, Lahore, shown in appendix figure A10, there are a mix of treatment and control assignments across schools within the same campus. We define the "closest school" as the school on the same campus as the teacher currently works, with grade levels closest to the teacher's current assignment. For example, for a first-grade primary school teacher, the "closest school" is the pre-primary school (nursery through kindergarten) on the same campus. However, for a fifth-grade primary school teacher, the "closest school" is the middle school (grades 6-8) on the same campus.

Our main specification is:

$$\begin{aligned} VA_{i,t-1} = \beta_0 &+ \beta_1 WorkatPP_i + \beta_2 Post_i + \beta_3 WorkatPP_i * Post_i + \beta_4 OppTreat_i \\ &+ \beta_5 OppTreat_i * Post + \beta_6 OppTreat_i * WorkatPP_i \\ &+ \beta_7 OppTreat_i * WorkatPP_i * Post + \chi_j + \epsilon_i \end{aligned} \tag{12}$$

This is similar to eq. 10 but adds in interaction with $OppTreat_i$, which is a dummy for whether the closest school is assigned the opposite treatment as the teacher's own school. The coefficient of interest is β_7, which tells us the difference in the extent of positive sorting for teachers who would face smaller switching costs to receive their ideal contract.

We find that when teachers' closest school is assigned the opposite treatment, there is a higher rate of positive sorting. Table 8 presents these results. Column 1 shows the extent of positive sorting for the full sample. Column 2 and 3 split the sample by whether the closest school received the same or the opposite treatment as the teacher's own school. The magnitude of positive sorting is about four times larger (0.027 SD versus 0.006 SD). Column 4 presents eq. 12. While there is a large difference in the extent of sorting, we cannot reject equality of the coefficients at the 10% level.

Another approach to test whether switching costs dampen the extent of positive sorting is to compare the contract choice versus the job choice in the second year. We can think of the contract choice decision as zero switching cost because teachers could remain at their current position but receive their preferred contract. Job choice decisions in the second year is a relatively high switching cost, as teachers move across schools in response to a short-term acquisition of their preferred contract. Comparing these two settings, we see substantial differences in the extent of positive sorting (0.03 SD versus 0.015 SD).

8 Potential negative consequences of sorting

8.1 Does performance pay attract "cheating" teachers?

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

First, we do not find that teachers who prefer performance pay are more likely to engage in distortionary teaching practices. They are significantly less likely to exhibit these behaviors than teachers who did not want performance pay. Figure 13 and appendix table A7 presents the treatment effects of objective performance pay along several dimensions of teaching pedagogy (classroom climate, differentiation, student-centered focus, and time spent on test preparation). The coefficient of interest is *Chose Perf Pay* Perf Pay Treat*, which tells us the heterogeneity in treatment effect by whether the teacher chose performance pay at baseline. The row titled $\beta(Treat + Treat * ChosePP)$ also presents the effect of performance pay for teachers who chose it. As we show in a companion paper (Andrabi and Brown, 2020), we find that objective performance pay results in a more negative classroom climate (more yelling, stricter discipline), more teacher-led time (less student-centered), and more time teaching to the test. However, these negative effects are almost completely concentrated among teachers who did not want performance pay. The overall effect of objective performance pay on classroom pedagogy rating is -0.41 SD for teachers who did not want performance pay as opposed to 0.16 SD for teachers who did want performance pay.

Second, we do not find that teachers who prefer performance pay ignore other areas of student development in order to maximize their pay. Figure 14 and appendix table A8 present results. At endline, we measure student satisfaction and socio-emotional development along five dimensions (survey items shown in appendix table C1). The effect of objective performance pay for teachers who chose flat pay is generally small and mixed across different dimensions. However, for teachers who chose performance pay, we find a significant positive effect on three of the five areas with an overall effect of 0.12 SD.

Finally, we can zoom in on different question types from the endline exam to see if treatment effects are concentrated among memorization-type questions, at the cost of other knowledge and skills. Table A6 column 1 presents the results for all question types. Column 2 presents results for questions that were pulled from external sources (PISA, TIMSS, and LEAPS), and hence were unlikely to be questions students would have been able to memorize. Columns 3 and 4 include questions from one grade below and one grade above the student's current year. We find significant effects of performance pay for teachers who chose it along all three areas, ranging from 0.11 SD to 0.20 SD. Combined, this evidence shows that the negative consequences that are often associated with performance pay are concentrated among teachers who did not want those contracts, not those who would sort in.

8.2 Does performance pay push out altruistic teachers?

Another concern is that performance pay may drive away teachers who are intrinsically motivated or pro-social. To test this, we measure teachers' pro-sociality, efficacy, competitiveness and time spent on school public goods (such as helping other teachers or assisting with extra-curriculars).[24] Figure 10 presents the difference along each characteristic for teachers who chose performance pay versus flat pay. We do not find that teachers who prefer performance pay spend significantly less time on providing public goods. Teachers who chose performance pay spend slightly more time on collaboration with other teachers and the same amount of time on administrative tasks. They do, however, spend less time meeting with parents and more time grading than those who chose flat pay. Teachers who prefer performance pay have similar levels of pro-sociality (as measured by signing up to volunteer to help financially disadvantaged students). They also are less likely to view their current job as a stepping stone to another job. This evidence suggests that performance pay does not attract significantly less altruistic teachers.

9 Policy Counterfactuals

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

We augment the simple framework from section 3.1 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 in a given period based on: i). the expected flow of wages, w_{jt}, for their remaining time in the labor force, τ, ii). the cost to change sectors if the job is not in the sector the employee currently works in, iii). non-wage utility, which is employee-job (ϵ_{ij}) and employee-job-time (ϵ_{ijt}) specific. Flat pay jobs pay a wage of 0, and performance pay jobs pay the piece-rate, p, times workers' priors about their output ($\hat{\theta} + \hat{\beta}$). Whether a job offers performance pay in a given year is denoted by δ_{jt}. Employees have full information about what contracts will be provided by each job over the length of their time in the labor force.

[24] Survey item description and sources are presented in appendix table C2. 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.

Employees choose which job has the highest predicted utility:

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

Table 9 presents the key parameter values used. To calculate the mean and standard deviation of teacher ability and behavioral effect of incentives, we make the following assumptions about the test score function. For the pre-period (and control group): $y_{it} = \theta_i + e_{it}$. For the treatment group during the intervention: $y_{it} = \theta_i + \beta_i + e_{it}$. We use our calculation of value-added in a given year for y_{it} and assume $Cov(e_{it}, e_{it+1}) = 0$. Here $t-1$ is one year before the intervention, t is the baseline and $t+1$ is the intervention year. The first and second moments of θ and β and their covariance are:

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

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

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

Finally, the accuracy of teachers' priors about their ability, α_θ, and behavioral response, α_β for existing teachers come directly from the contract choice experiment. We use a separate set of lower accuracy, but non-zero, priors for individuals who are not currently teachers. The values chosen take into account evidence from this study across teacher tenure and evidence on applicant teacher accuracy from Leaver et al. (2019) and Johnston (2020). We also include optimistic values of the parameters to take into account that longer term policies would likely result in better understanding of the performance metrics used.

We find that the introduction of a long term performance pay contract induces a fair amount of sorting, though effects vary depending on the use of pessimistic versus optimistic parameter values. Figure 15 presents the effects over time of introducing a 1 year, 10 year or 30 year performance pay policy. The effect of a 1 year policy is just the average behavioral response (0.07 SD). Under a 10 year policy, there is an average effect of 0.075 SD (0.10 SD) if using pessimistic (optimistic) parameter values during the time the policy is in place. Under optimistic parameters, there are also effects after the policy is removed due to the attraction of higher performing teachers that then stay in the profession even after the policy is removed. The introduction of a 30 year policy results in an average effect of 0.09 SD (0.17 SD) under pessimistic (optimistic) parameters. These effects are 1.3-2.4x larger than the one year effects of performance pay.

10 Conclusion

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

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

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

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

11 References

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 2020, 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 Kingdon, "Assessing Teacher Quality in India," *Working Paper*, October 2014, p. 31.

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, "Unions, Salaries, and the Market for Teachers: Evidence from Wisconsin," *SSRN Electronic Journal*, 2017.

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.

Bruhn, Miriam and David McKenzie, "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," October 2009, *1* (4), 34.

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.

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.

Lavy, Victor, "Using Performance-Based Pay to Improve the Quality of Teachers," *The Future of Children*, 2007, *17* (1,), 87–109.

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," *RISE Programme Working Paper*, June 2019, p. 41.

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 Criminal Behavior," *Working Paper*, May 2019, p. 63.

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.

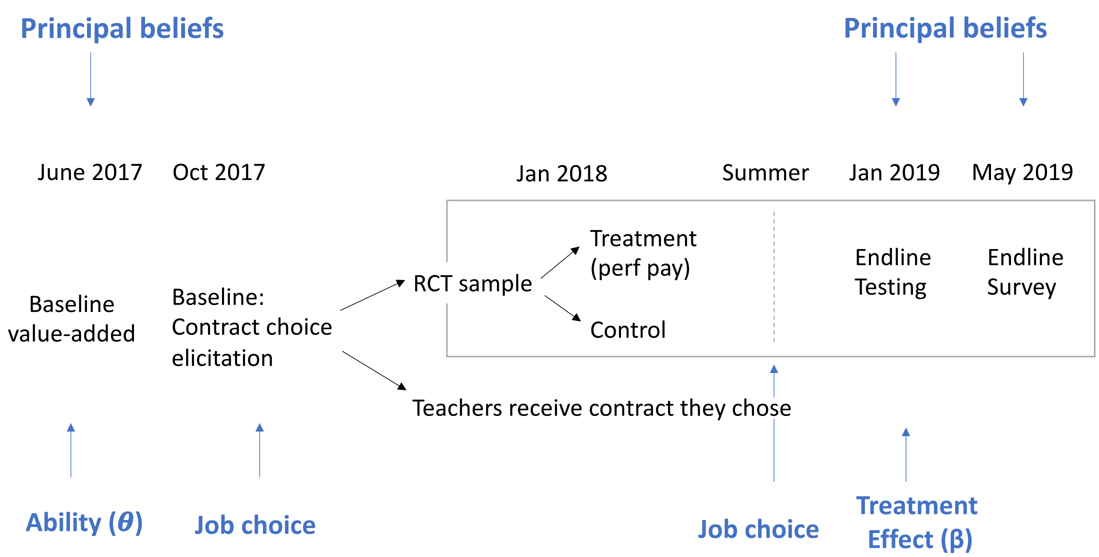
World Bank, *Systems Approach for Better Education Results (SABER)* 2018.

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

_ , *Enterprise Surveys* 2019.

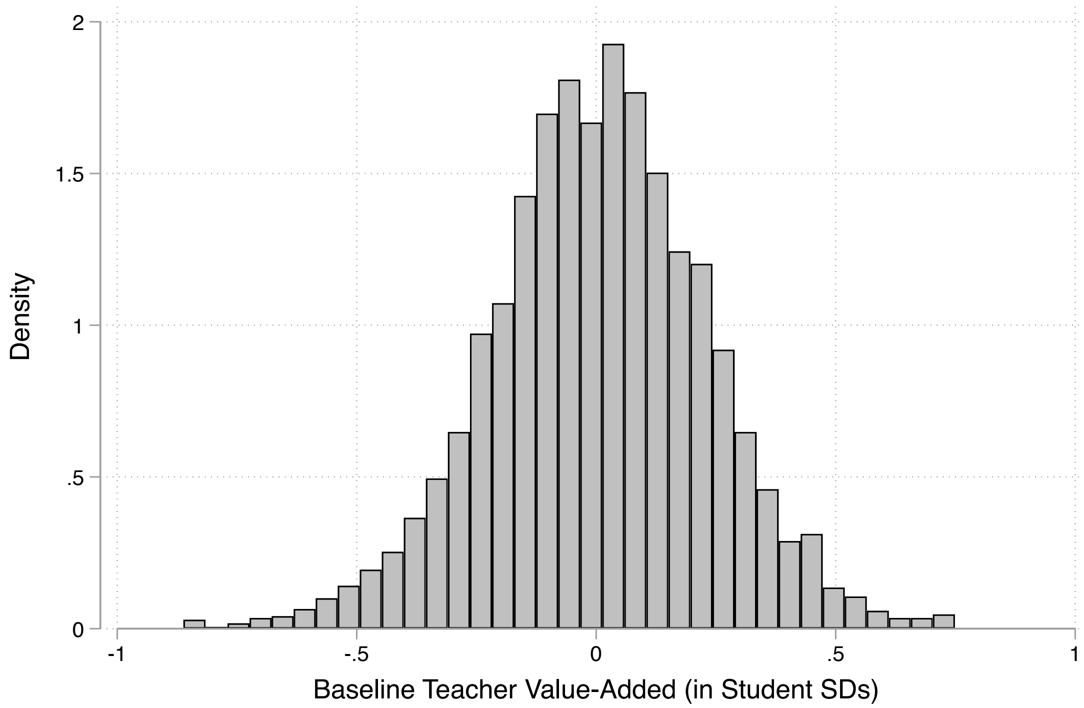
12 Figures

Figure 1: Experiment Timeline



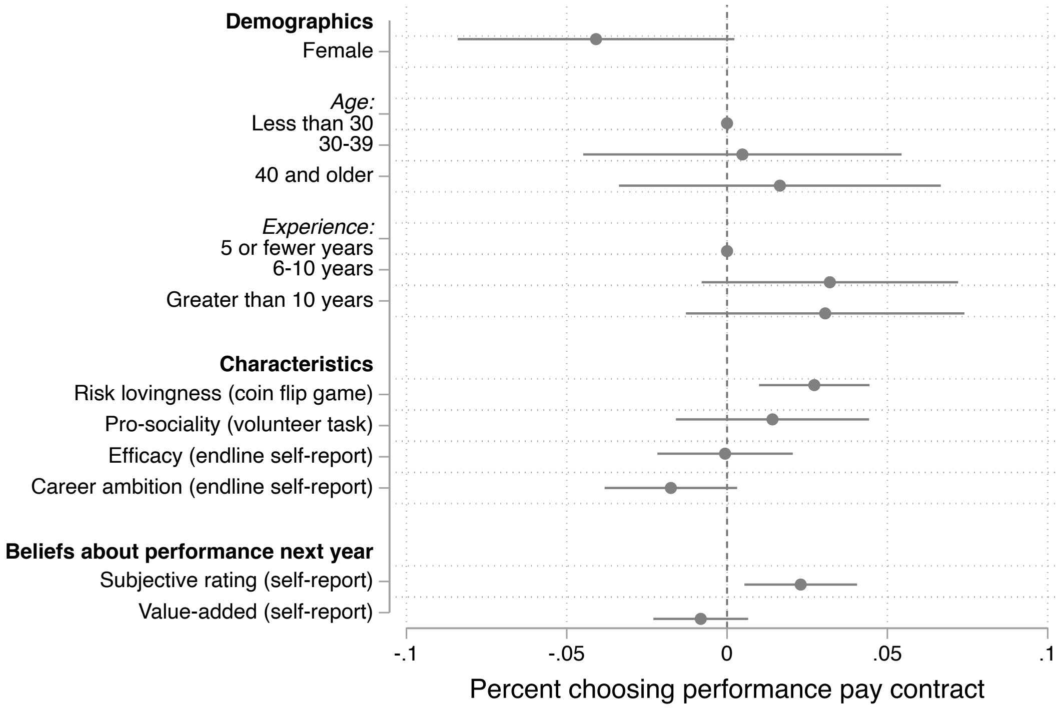
Notes: The figure presents the experimental timeline from June 2017 through May 2019. Our measure of ability comes from the calculation of teacher value-added in June 2017 prior to the introduction of the treatments. Our measure of the behavioral effect of performance pay comes from comparing the treatment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

Figure 2: Distribution of Teacher Value-Added at Baseline



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

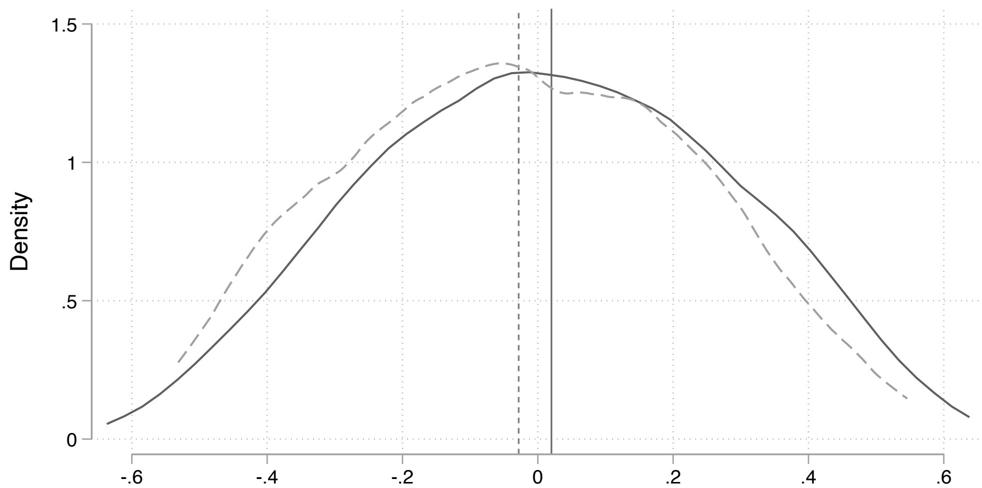
Figure 3: Predictors of contract choice



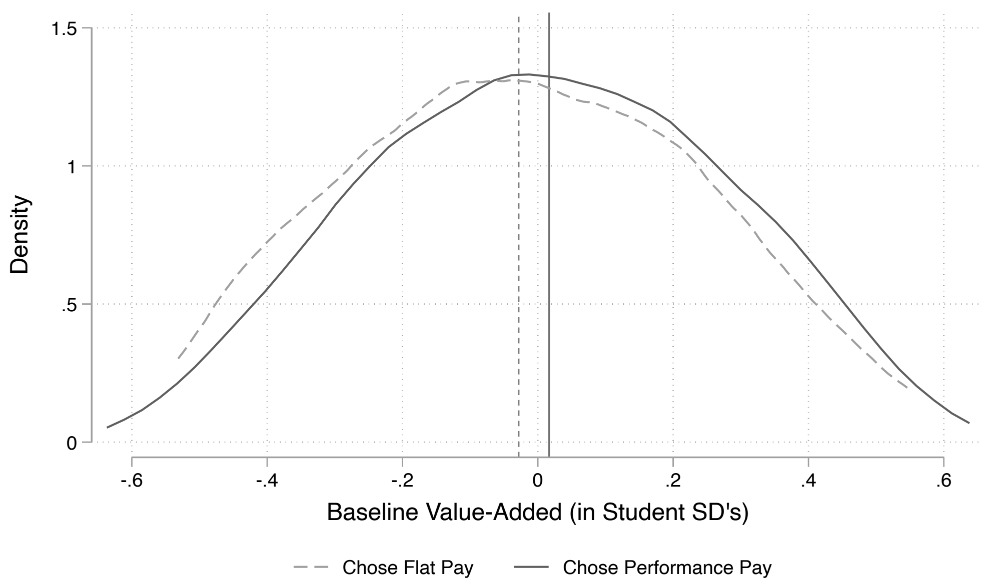
Notes: This figure presents coefficients and 95% confidence intervals of bivariate regressions of teacher's contract choice on teacher demographics, characteristics and beliefs. Teacher's contract choice is a dummy for whether they selected a performance pay or flat pay contract. All independent variables,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

Figure 4: Distribution of Baseline Value-Added by Contract Choice

Panel A: Objective Performance Metric



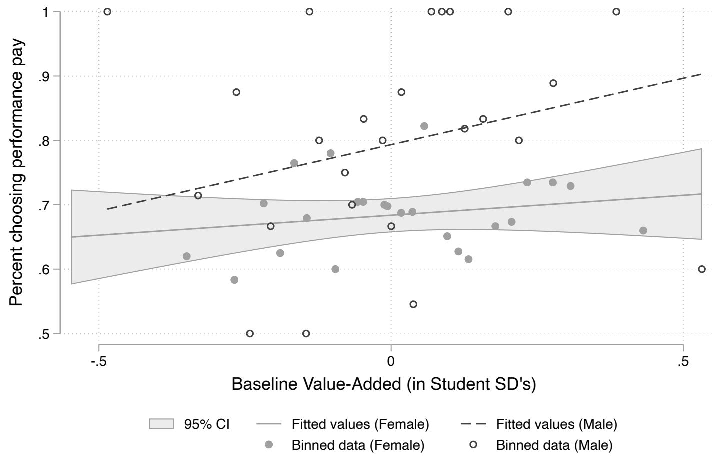
Panel B: Subjective Performance Metric



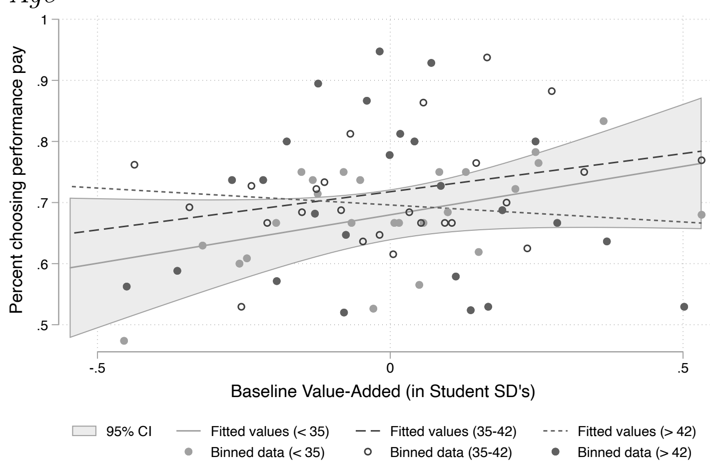
Notes: This figure plots the distribution of baseline teacher value-added for teachers who chose performance pay (solid line) versus flat pay (dotted line). Panel A presents results for the choice between objective (value-added based) performance pay versus flat pay. Panel B presents results for the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

Figure 5: Relationship between Value-Added and Contract Choice by Demographics

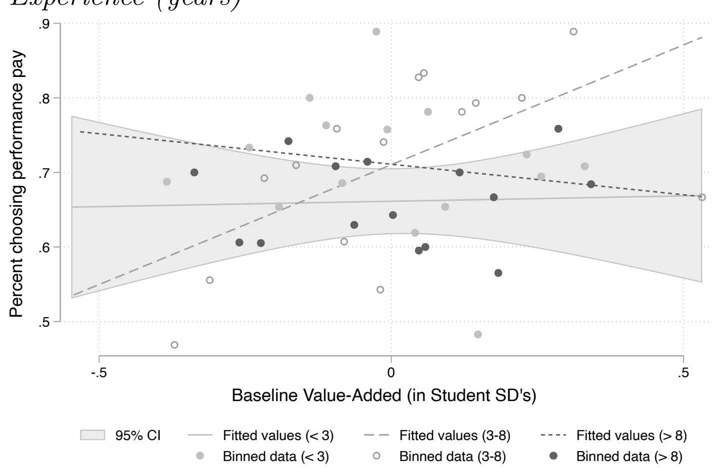
Panel A: By Teacher Gender



Panel B: By Teacher Age



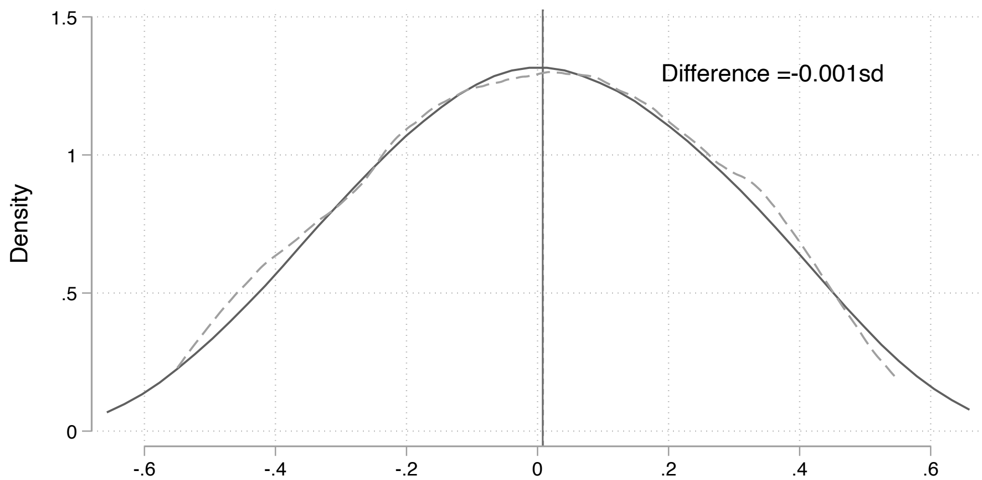
Panel C: By Teacher Experience (years)



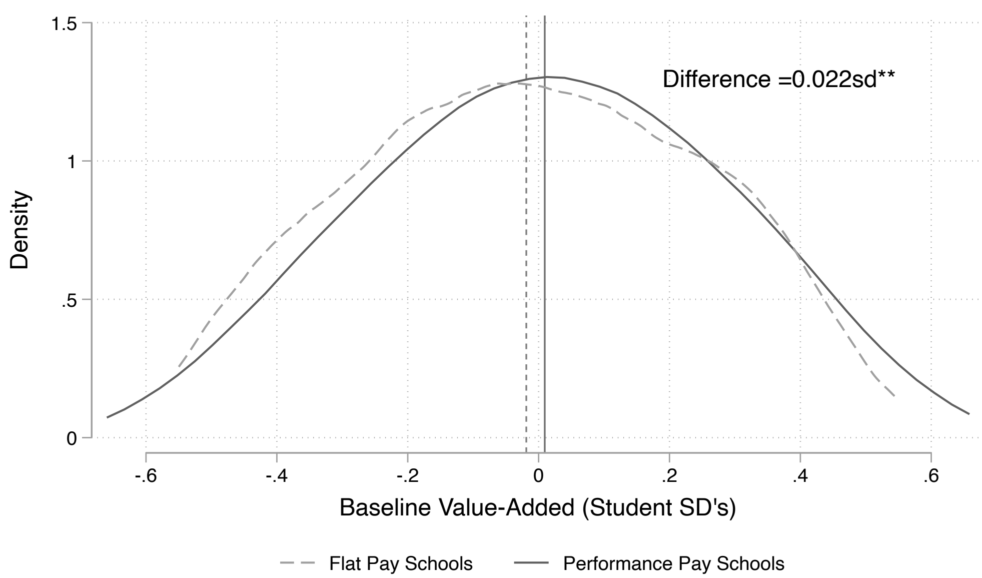
Notes: These figures plot the relationship between teacher quality as measured by baseline value-added and teacher's contract choice. The graph plots binned values of *Teacher Baseline Value-Added* by the percent of teachers in that bin that chose performance pay. Results are shown by teacher characteristic.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

Figure 6: Distribution of Teacher Baseline Value-Added by School and Year

Panel A: December 2017 (Baseline)

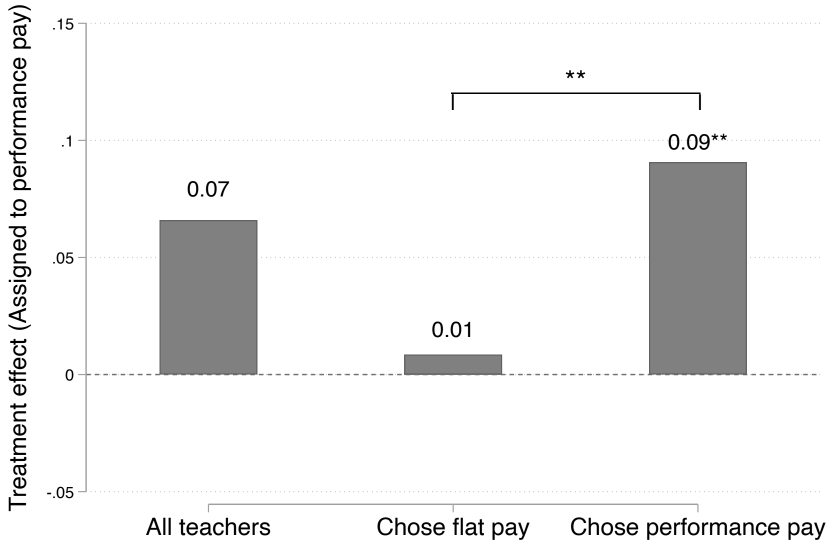


Panel B: December 2018 (One year after treatment announcement)



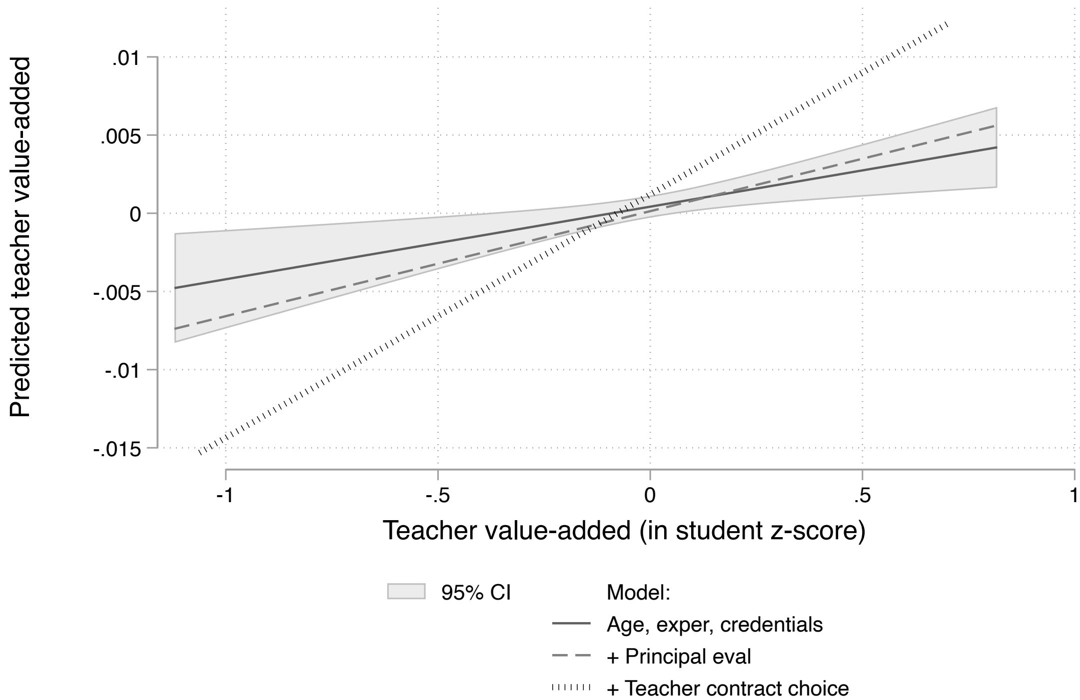
Notes: These figures plots the distribution of baseline teacher value-added for teachers in performance pay versus flat pay schools. Panel A provides the distribution in December 2017 (one month before the treatments are announced). Panel B provides the distribution in December 2018 (11 months after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

Figure 7: Treatment Effect by Contract Choice



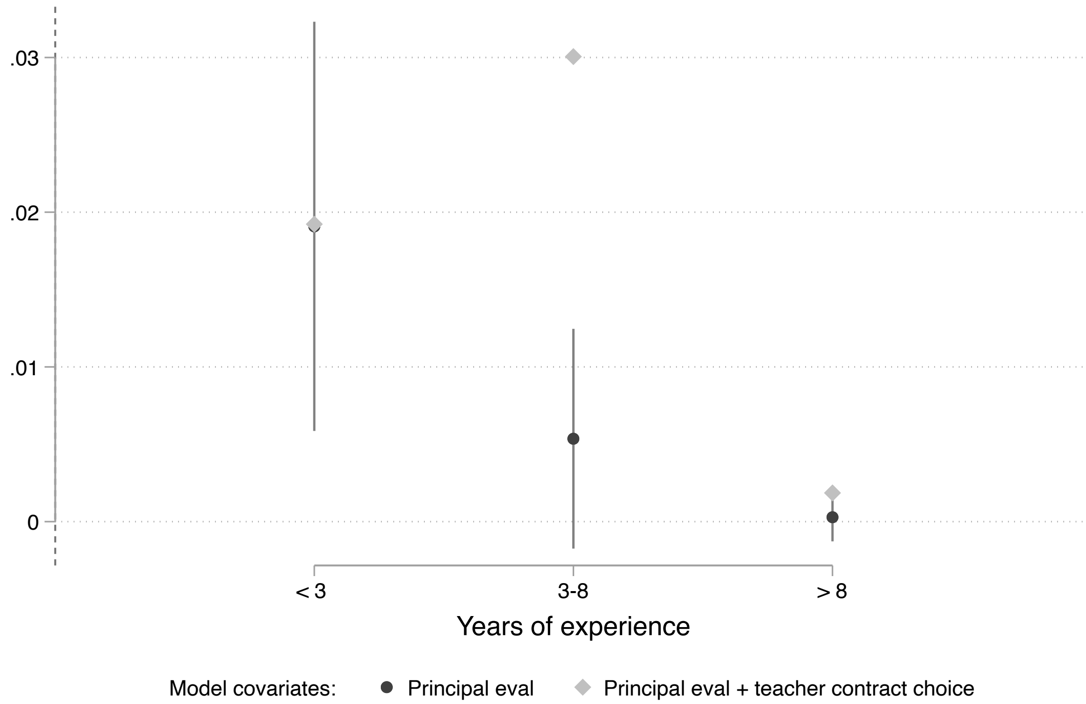
Notes: This figure presents the treatment effects from the performance pay on endline test scores. The first bar presents the effects for all teachers. The second bar presents the treatment effects for teachers who stated in the baseline contract choice exercise that they wanted a flat pay contract.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

Figure 8: Predicting Teacher Value-Added



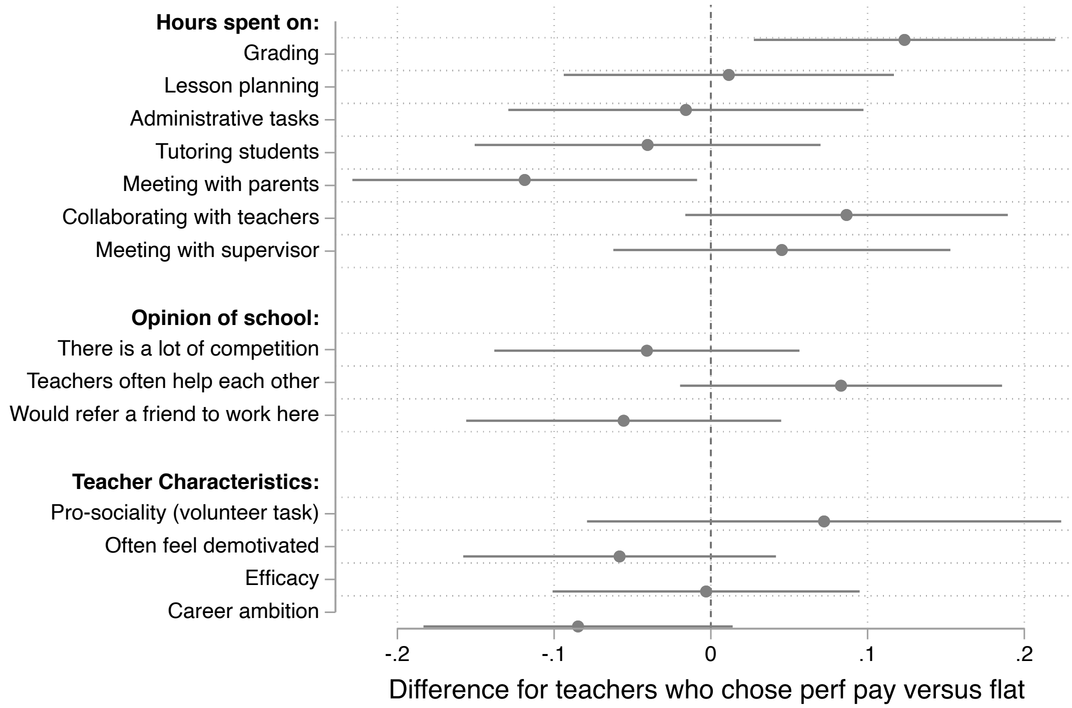
Notes: This figure presents the relationship between value-added and predicted value-added for three different models. The first model (solid line) just includes teacher demographics (age, experience and credential-type fixed effects). The second model (dashed line) uses demographics and principal evaluation.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

Figure 9: Predicting Teacher Value-Added by Experience



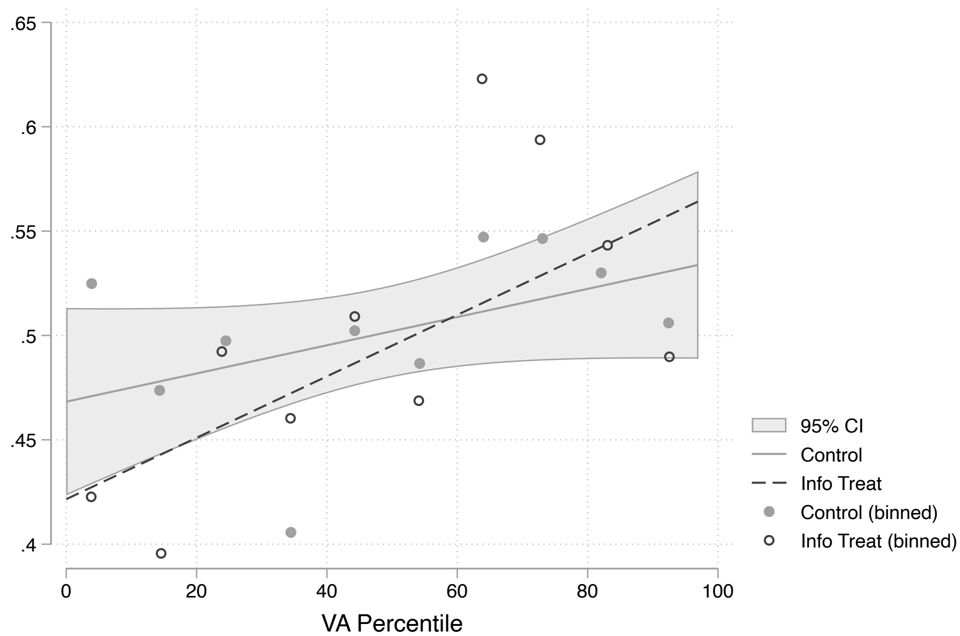
Notes: This figure presents the coefficient and 95% confidence intervals for predicted value-added on value-added for two different models. The first model (black circle) uses principal evaluation. The second (gray diamond) model includes principal evaluation and teacher's baseline contract choice. Results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

Figure 10: Predictors of contract choice



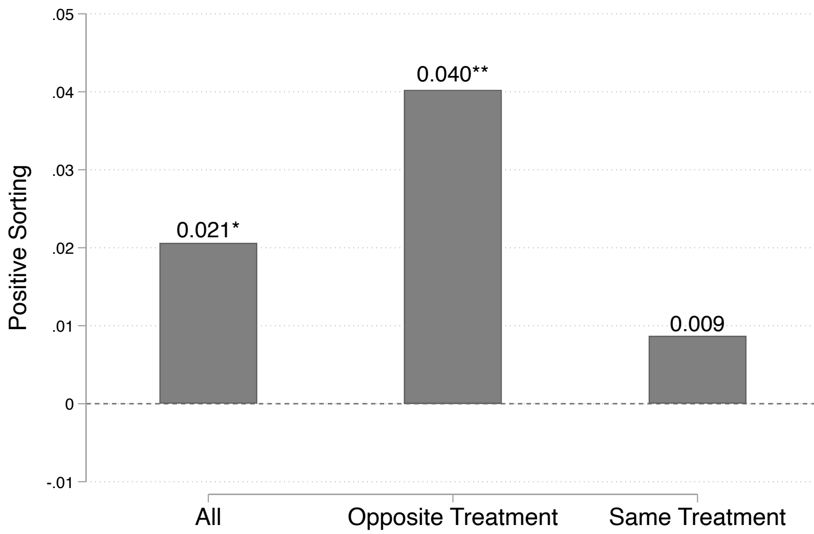
Notes: This figure presents coefficients and 95% confidence intervals of bivariate regressions of teacher time use and characteristics on teacher's contract choice on. Teacher's contract choice is a dummy for whether they selected a performance pay or flat pay contract. All outcomes are standardized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

Figure 11: Beliefs and Contract Choice by Teacher Value-Added



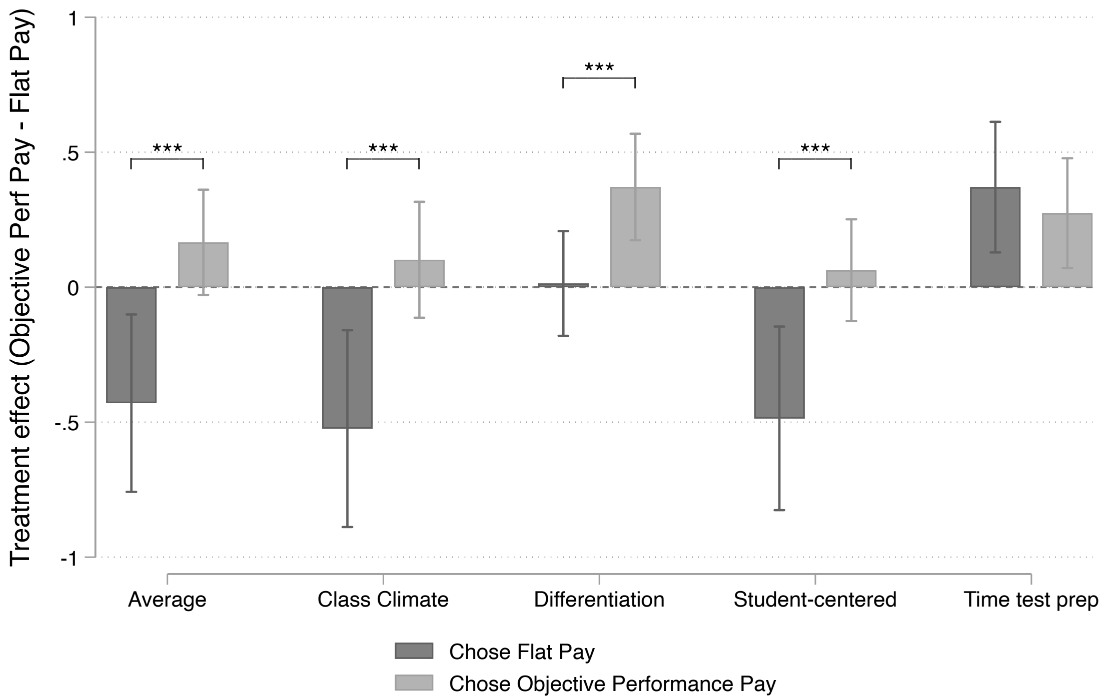
Notes: The figure shows the relationship between teacher's contract choice and their value-added. The solid line, 95% confidence interval, and circles present the relationships for control teachers. The dotted line and white circles show the relationship for teachers who received information about their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

Figure 12: Positive Sorting by Closest School's Treatment



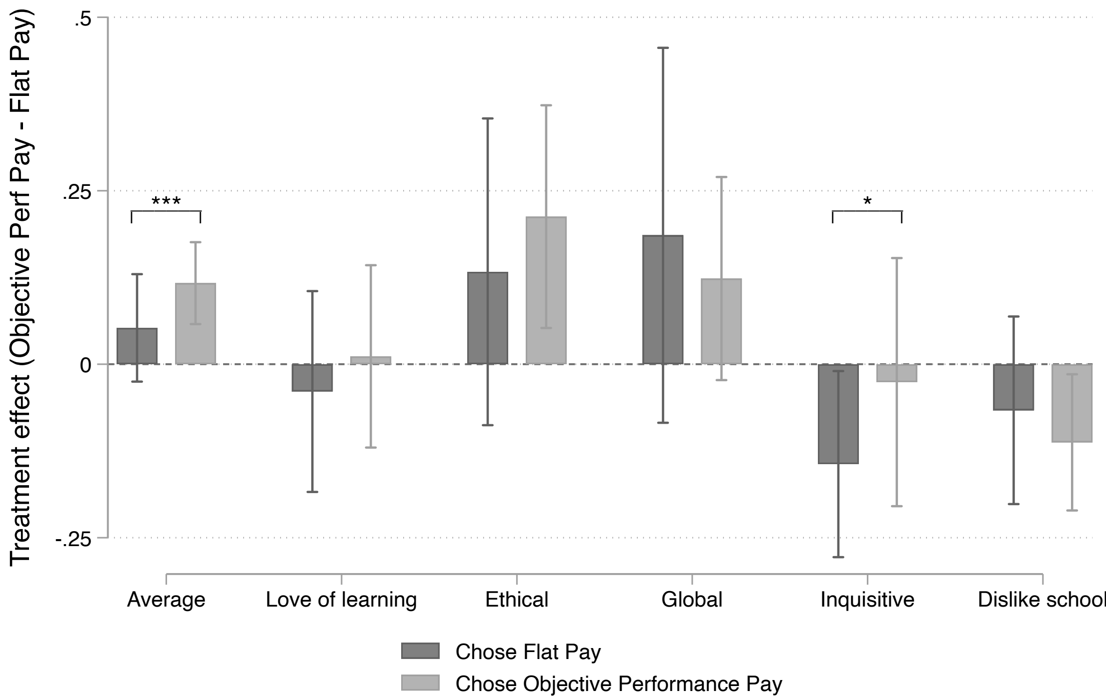
Notes: This figure presents the difference in baseline value-added among teachers employed at performance pay versus flat pay schools at endline. The first bar presents the results for all teachers. The second presents the results for teachers whose closest school to them was assigned the opposite treatment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

Figure 13: Treatment Effects on Classroom Observations by Contract Choice



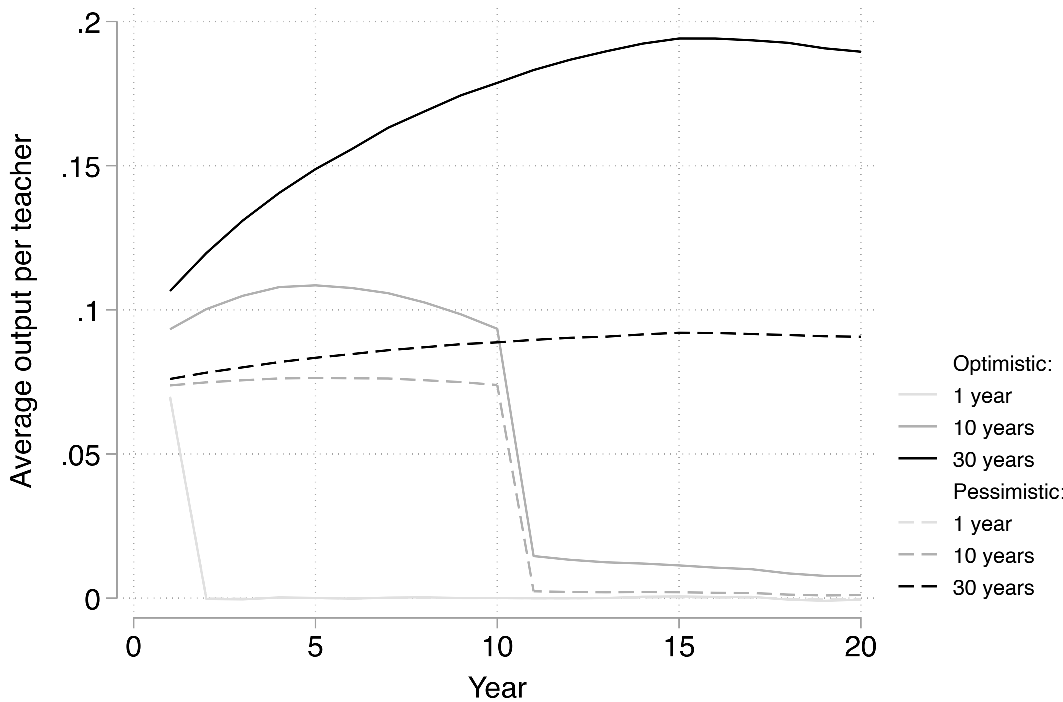
Notes: This figure presents the treatment effect and 95% confidence intervals of objective performance pay relative to flat pay for teachers who chose flat pay (left bar) versus chose performance pay (right bar). Outcomes are from classroom observation data. Standard errors are clustered at the school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

Figure 14: Treatment Effects on Student Surveys by Contract Choice



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

Figure 15: Policy Simulations



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

13 Tables

Table 1: Descriptive Statistics about Study Sample and Comparison Sample

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

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

Table 2: Teacher Value-Added by Contract Choice

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

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

Table 3: Teacher Value-Added by Contract Choice and Demographics

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

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

Table 4: Teacher Quality by School

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

Notes: This table presents the relationship between teacher quality (as measured by teacher value-added) and where teachers choose to work. The outcome is *Teacher Baseline Value-Added*, measured using test score data from the two years prior to the intervention. *Performance Pay School* is a dummy for if a teacher works at a school that is assigned to a performance pay treatment contract (as compared to works at a school which was assigned a control flat pay contract). *Post* is a dummy that is equal to 0 in December 2017 and 1 in December 2018. Data is at the teacher-year level. Column (1) presents basic specification (eq. 10). Columns (2)-(4) add additional controls. Standard errors are clustered at the school level. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table 5: Treatment Effect by Contract Choice

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

Notes: This table presents the treatment effect of performance pay contracts on endline test scores by teacher characteristics. The outcome is students' standardized z-score from the endline test conducted in January 2019. *Assigned Perf Pay Treat* is a dummy for whether a teacher taught at a school assigned to performance pay at baseline. *Chose Perf Pay* is a dummy variable for whether a teacher chose objective performance pay or flat pay during the baseline choice exercise. *Principal Rating of Teacher* is the baseline subjective rating z-score of the teacher by their principal. Column (1) presents the treatment effect for all teachers. Column (2) 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 6: Principal Beliefs about Teacher Quality

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

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

Table 7: Teacher Value-Added by Contract Choice - Information Treatment

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

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

Table 8: Positive Sorting by Closest School's Treatment

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

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

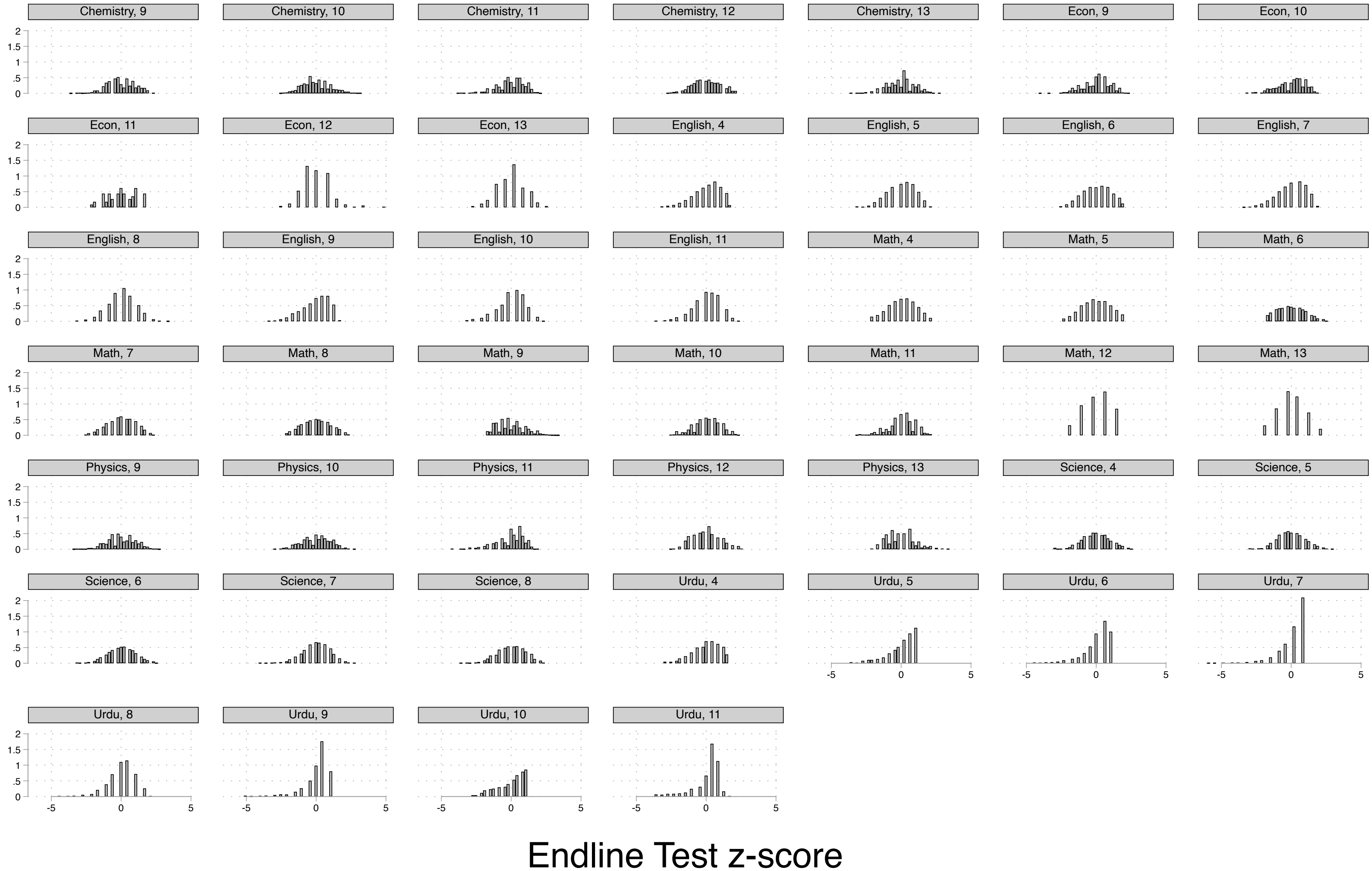
Table 9: Values of Key Parameters

Parameter	Value		Source/Calculation
	Pessimistic	Optimistic	
Mean ability, μ_θ	0	-	Test scores
SD ability, σ_θ	0.15	0.30	Test scores
Mean behavioral effect, μ_β	0.07	-	Test scores
SD behavioral effect, σ_β	0.14	0.28	Test scores
Correlation θ and β, $\rho_{\theta\beta}$	$-1.94\text{x}10^{-4}$	-	Test score
Fraction new entrants	0.3	-	Admin data
Job-employee specific utility, σ_ϵ	\$360	-	Admin data
Job-employee time shocks, σ_e	\$180	-	Admin data
Mean cost to change professions, μ_c	\$1,120	-	Survey
SD cost to change professions, σ_c	\$1,200	-	Survey
Accuracy of priors (existing teachers) θ, α_θ	0.049	0.075	Admin data/survey
Accuracy of priors (existing teachers) β, α_β	0.025	0.038	Admin data/survey
Accuracy of priors (non-teachers) θ, α_θ	0.033	0.050	Admin data/survey
Accuracy of priors (non-teachers) β, α_β	0.017	0.033	Admin data/survey

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

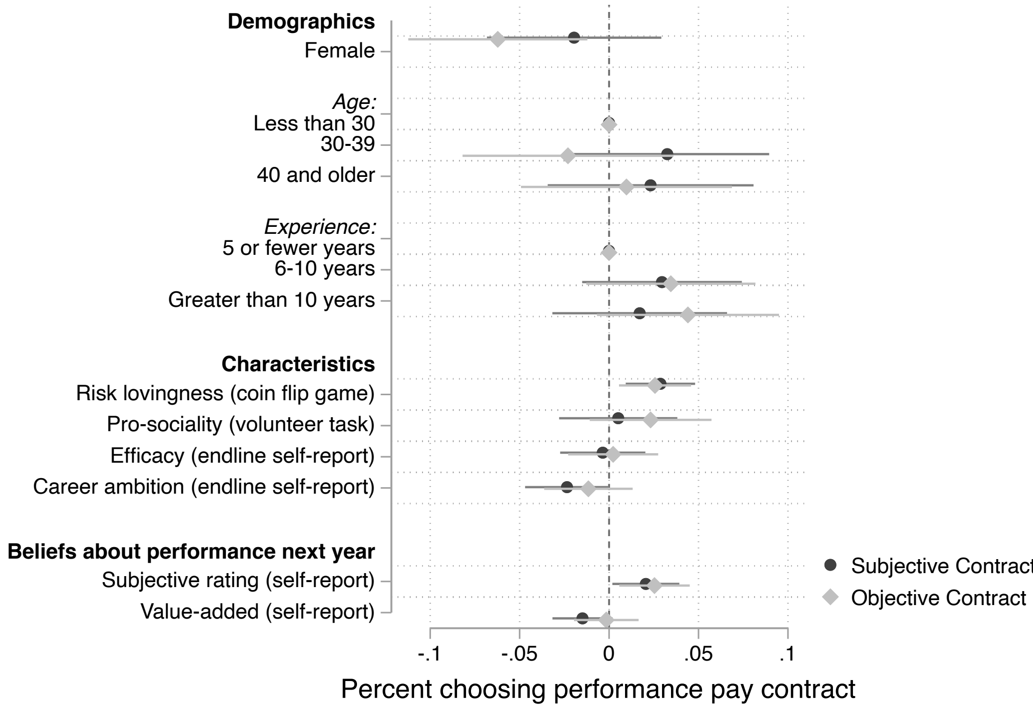
Appendix A - Supplementary Tables and Figures

Figure A1: Distribution of Endline Test Scores



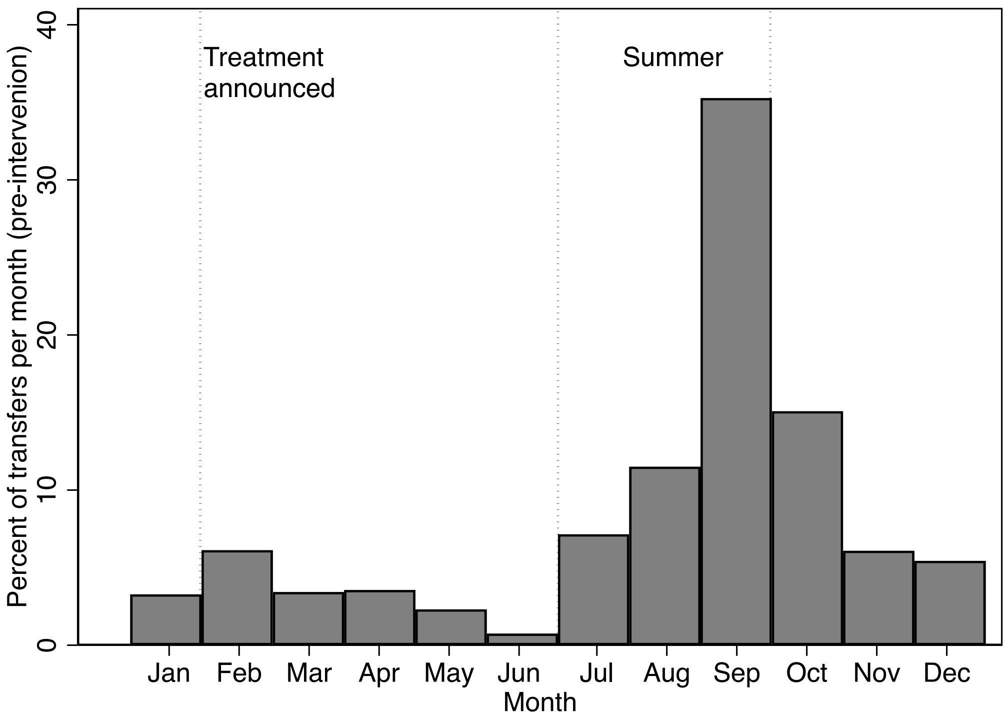
Notes: This figure presents the standardized distribution of student scores across each exam administered at endline. The endline test was conducted in January 2019 across grades 4-13 in English, Urdu, Math, Science and Economics. In grades 9-13, students took the science exam in the class they were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

Figure A2: Predictors of contract choice



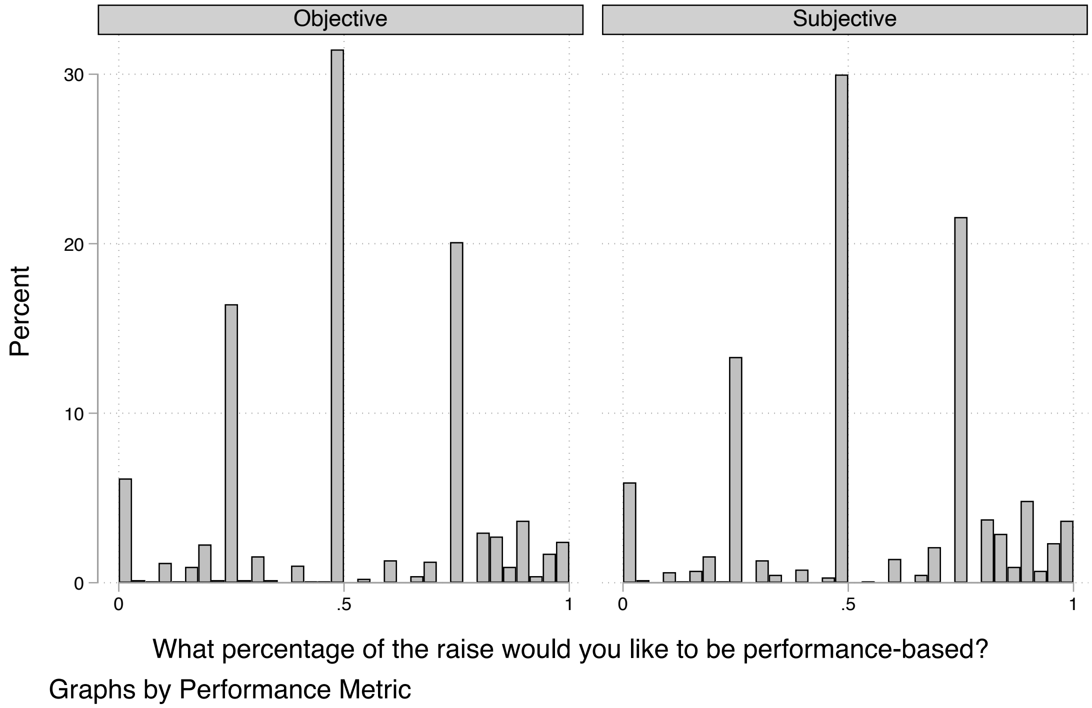
Notes: This figure presents coefficients of bi-variate regressions of teacher's contract choice on teacher demographics, characteristics and beliefs. Teacher's contract choice is a dummy for whether they selected a performance pay or flat pay contract. All independent variables other than gender, age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

Figure A3: Teacher transfers across campuses within school system



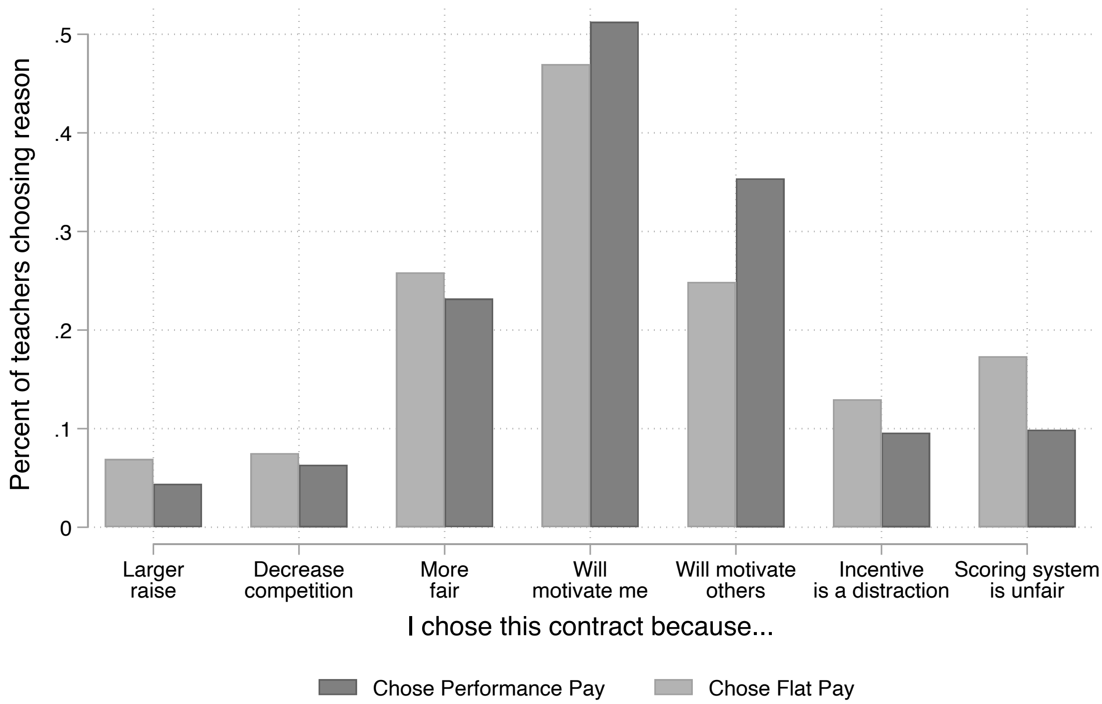
Notes: This figure plots the percent of transfers across schools within the system by month. Transfer data is from administrative schools records from 2015, prior to the intervention.

Figure A4: Distribution of contract choice by performance metric



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

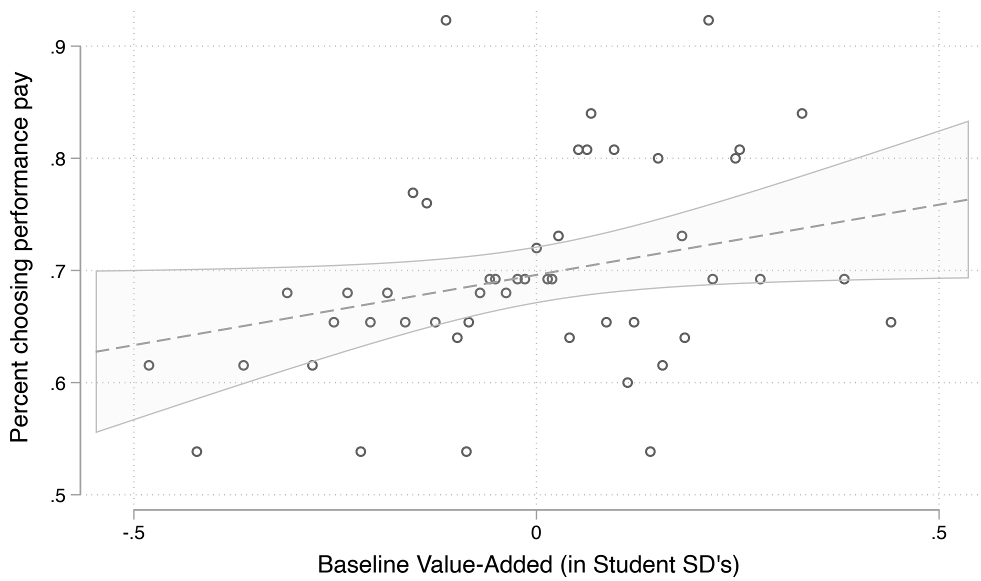
Figure A5: Teachers stated reasons for selecting performance pay or flat pay contract



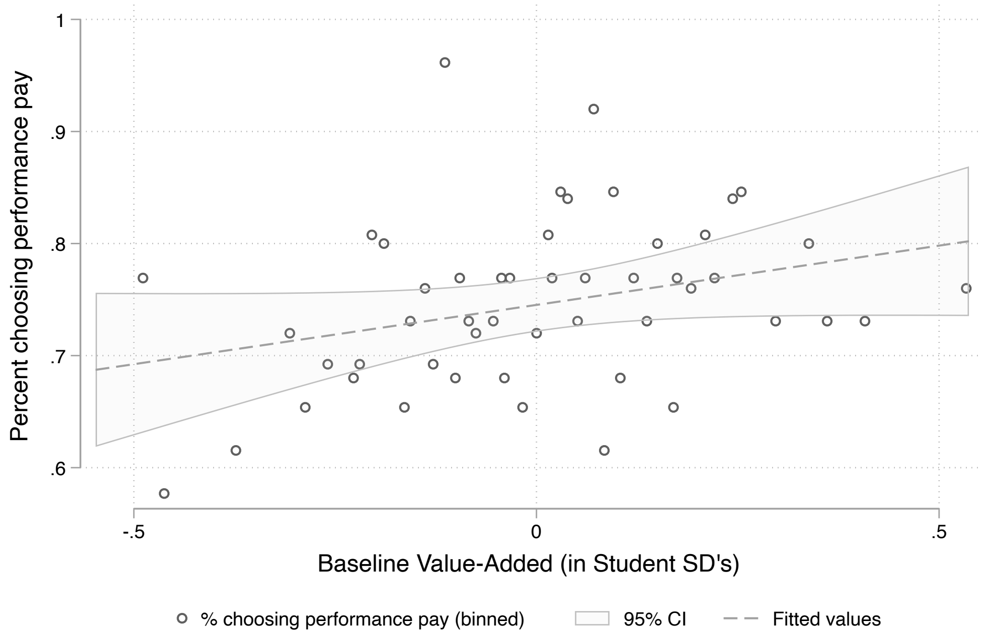
Notes: This figure plots teachers responses to the question *Why did you select this contract?*. The graph shows the percent of teachers that selected each reason. Teachers are allows to select multiple reasons, if applicable. The light gray bars plot responses for teachers who chose a flat pay contract.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

Figure A6: Relationship between Value-Added and Contract Choice

Panel A: Objective Performance Metric



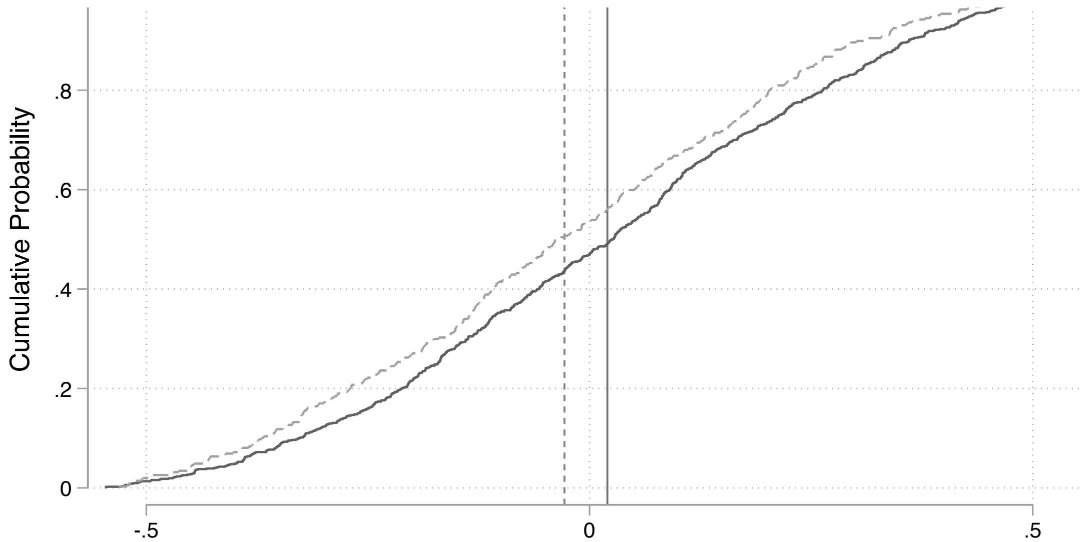
Panel B: Subjective Performance Metric



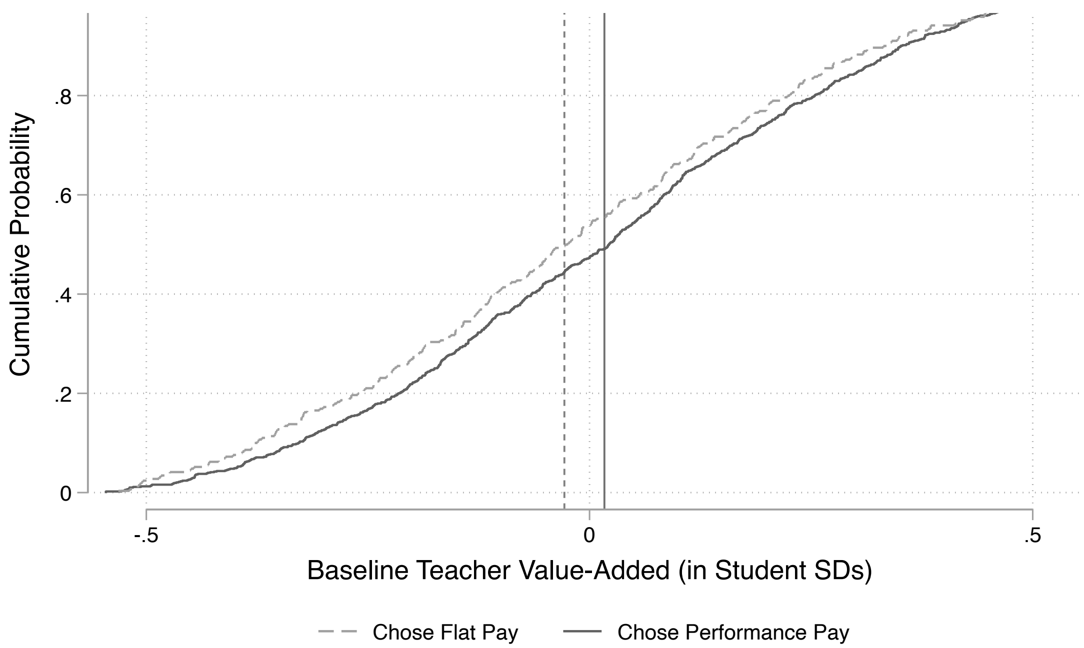
Notes: These figures plot the relationship between teacher quality as measured by baseline value-added and teachers contract choice. The graph plots binned values of *Teacher Baseline Value-Added* by the percent of teachers in that bin that chose performance pay. Panel A presents results for the choice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

Figure A7: Cumulative Distribution Function of Baseline Value-Added by Contract Choice

Panel A: Objective Performance Metric



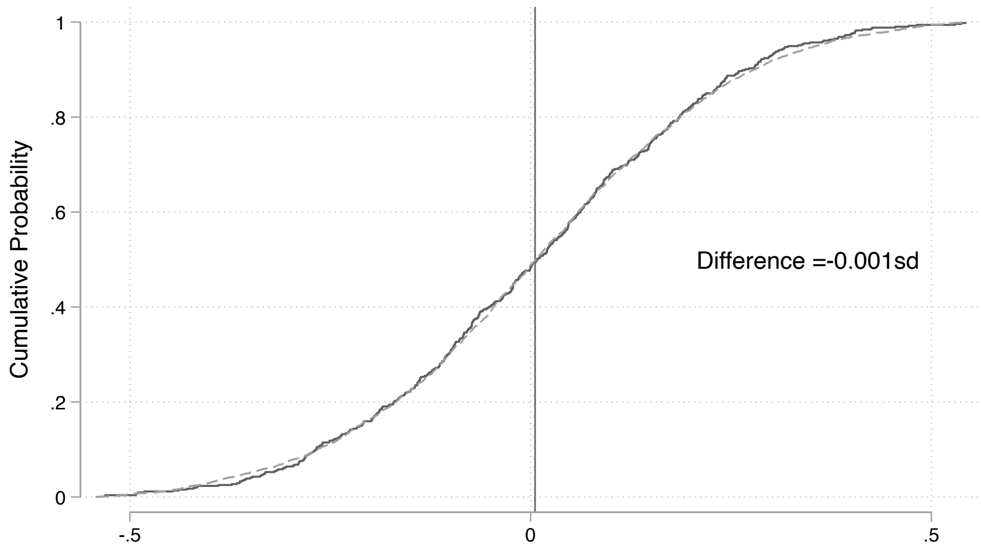
Panel B: Subjective Performance Metric



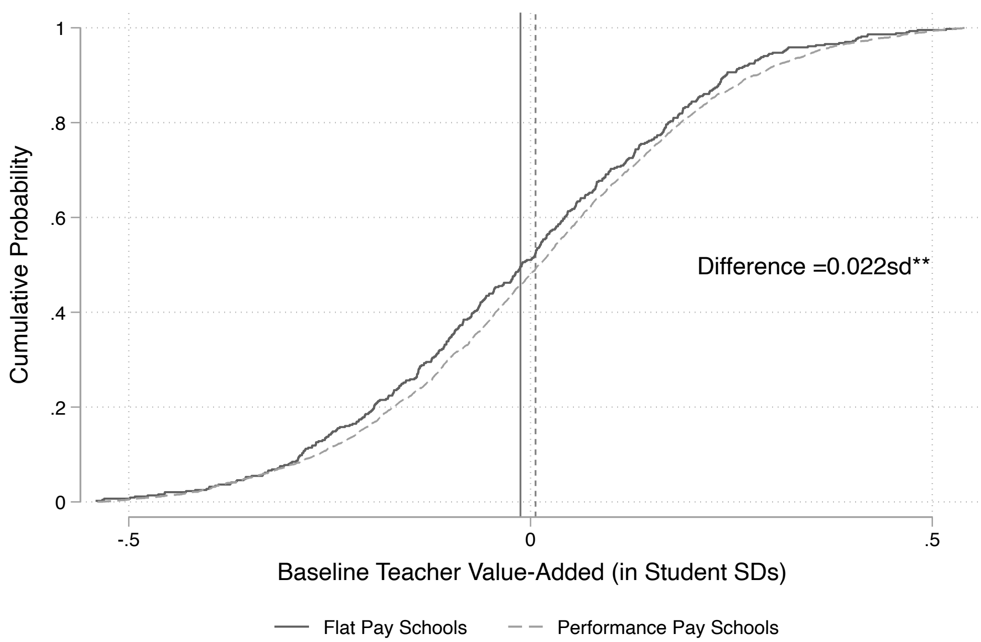
Notes: This figure plots the CDF of baseline teacher value-added for teachers who chose performance pay (solid line) versus flat pay (dotted line). Panel A presents results for the choice between objective (value-added based) performance pay versus flat pay. Panel B presents results for the choice between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

Figure A8: CDF of Teacher Baseline Value-Added by School Treatment and Year

Panel A: December 2017 (Baseline)

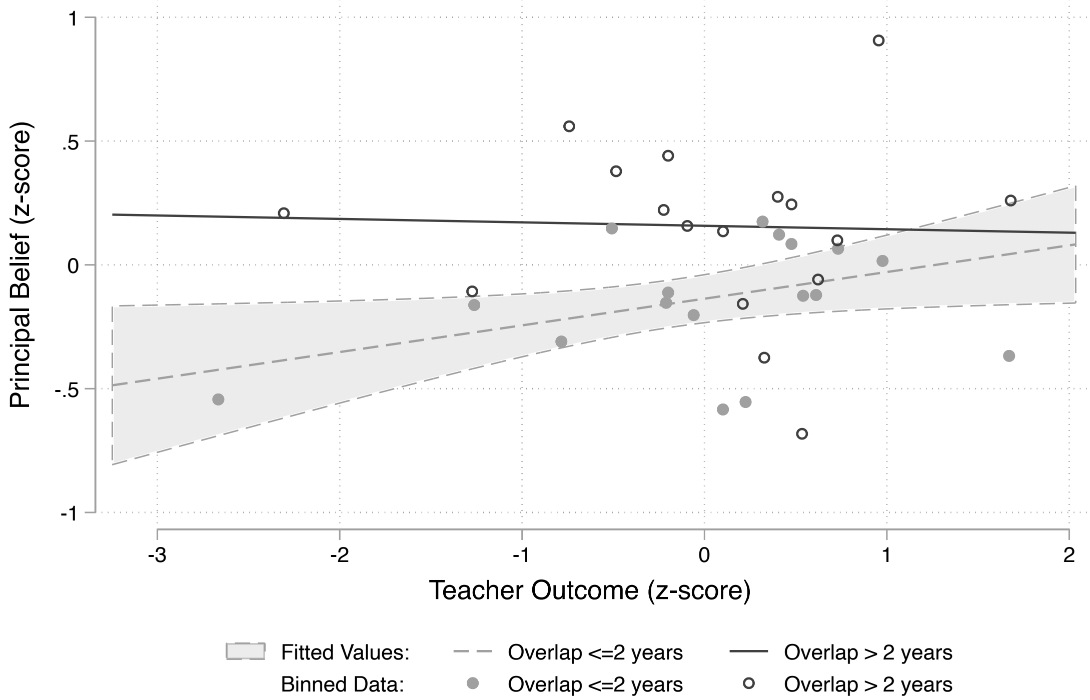


Panel B: December 2018 (One year after treatment announcement



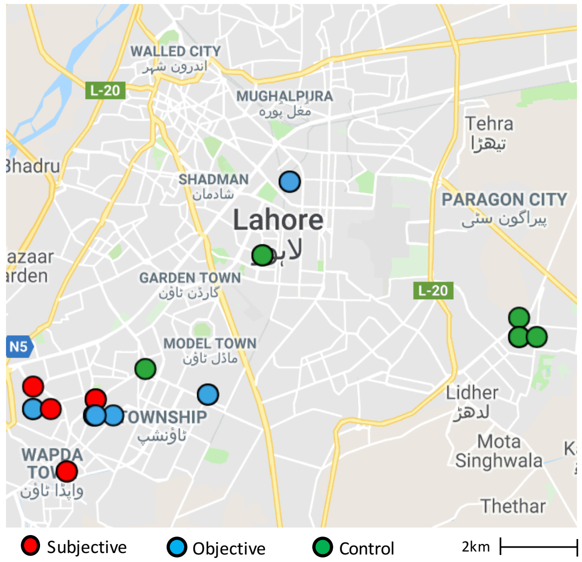
Notes: These figures plots the CDF of baseline teacher value-added for teachers in performance pay versus flat pay schools. Panel A provides the distribution in December 2017 (one month before the treatments are announced). Panel B provides the distribution in December 2018 (11 months after the treatments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

Figure A9: Principal Beliefs about Teacher Outcome by Overlap of Principal and Teacher



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

Figure A10: Treatment Distribution Map, Lahore



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

Table A1: Baseline Covariates

Variable	(1) Control N/ [Clusters]	(1) Control Mean/ SE	(2) Objective Treatment N/ [Clusters]	(2) Objective Treatment Mean/ SE	(3) Subjective Treatment N/ [Clusters]	(3) Subjective Treatment Mean/ SE	T-test Difference (1)-(2)	T-test Difference (1)-(3)	T-test Difference (2)-(3)
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 A2: Treatment Effect by Contract Choice

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

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

Table A3: Relationship between Teacher Value-Added and Characteristics

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

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

Table A4: Sorting Controlling for Teacher Characteristics

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

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

Table A5: Baseline Covariates - Neighboring School's Treatment

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

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

Table A6: Treatment Effect by Contract Choice, Across Question Type

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

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

Table A7: Treatment Effects on Classroom Observations by Contract Choice

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

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

Table A8: Treatment Effects on Student Survey by Contract Choice

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

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

Appendix B Proofs

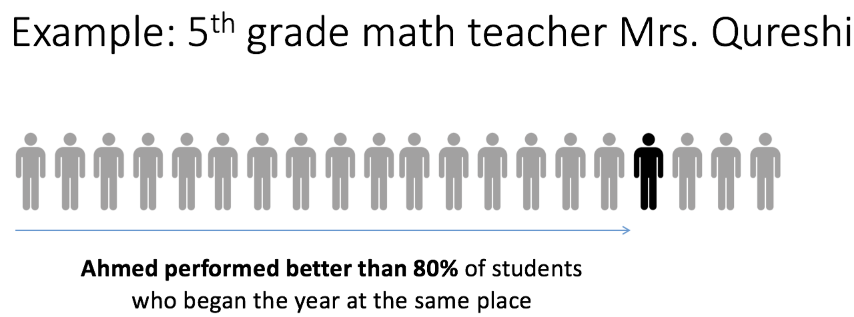
Proof of eq. 5

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

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

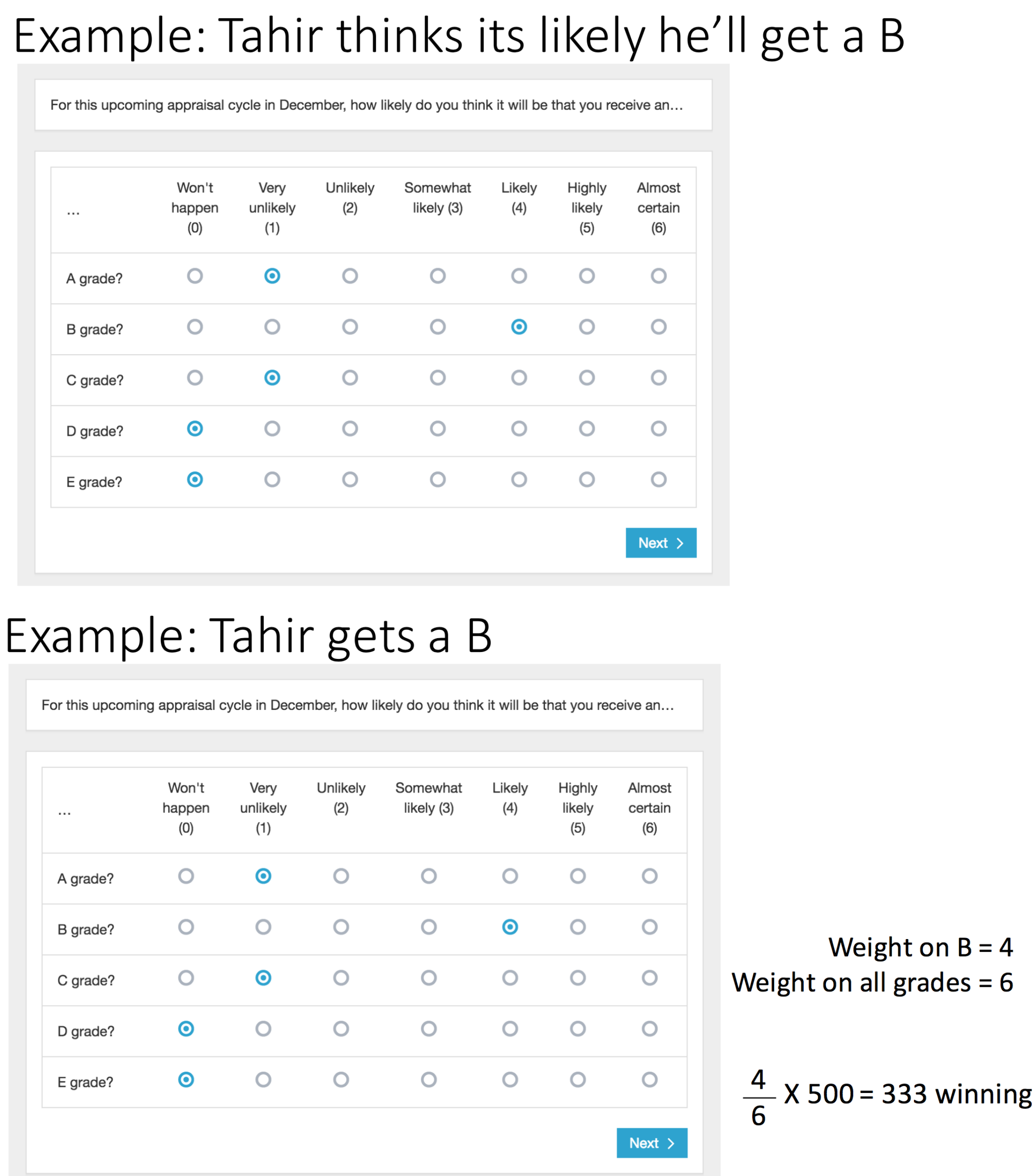
Appendix C Experimental Design Implementation

Figure C1: Screen capture from survey video: Calculation of percentile VA



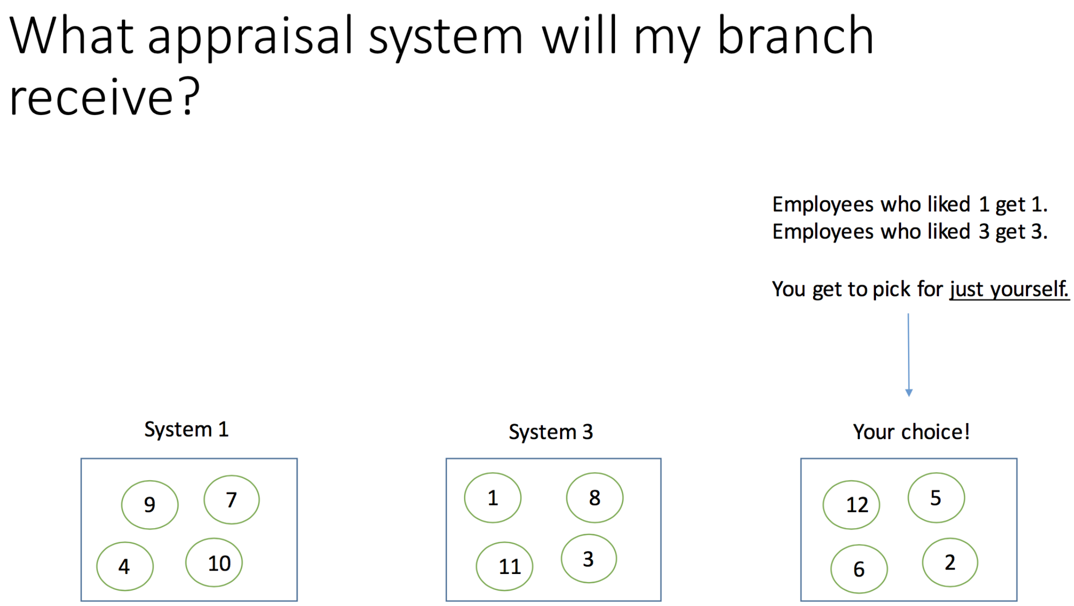
Notes: Screen capture from the video explaining to teachers how percentile value-added was calculated, giving teachers practical examples.

Figure C2: Screen capture from baseline survey: Incentivized belief distribution elicitation



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

Figure C3: Screen capture from baseline survey: Contract randomization



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

Figure C4: Example Performance Criteria

PERFORMANCE APPRAISAL - FORM D

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

Plan 1: Manager Appraisal of Effort

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

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

Figure C5: Example Midterm Information

Dear Emp - 2890 ,

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

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

These criteria are:

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

Your midterm performance is:

Unsatisfactory	Satisfactory	Good	Very Good	Excellent

Ok

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

Table C1: Socio-Emotional Outcomes Student Survey

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

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

Table C2: Teacher Characteristics - Survey Items

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

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