

Subjective versus Objective Incentives and Teacher Productivity

Tahir Andrabi and Christina Brown

Abstract

A central challenge facing firms is how to incentivize employees. While objective, output-based incentives may be theoretically ideal, in practice they may lead employees to reduce effort on non-incentivized outcomes and may fail in settings where effort is weakly tied to output. We study the effect of subjective incentives (manager performance evaluation) and objective incentives (test score-based) relative to no incentives for teachers using an RCT in 230 Pakistani schools. First, we show that subjective and objective incentives both increase test scores and have similar magnitude effects. However, objective incentives decrease non-test score student outcomes relative to subjective incentives. Second, we show that teachers' effort response is very different under each scheme, with attendance increasing under subjective and teaching quality decreasing under objective. Finally, we rationalize these effects through the lens of a moral hazard model with multi-tasking. We use within-treatment variation to isolate the causal effect of contract noise and distortion and show that these channels explain most of our reduced form effects.

Subjective versus Objective Incentives and Teacher Productivity

Tahir Andrabi
Pomona College
tandrabi@pomona.edu

Christina Brown
University of Chicago
christinabrown@uchicago.edu

Acknowledgements:

We are incredibly thankful for support and advice from Supreet Kaur, Christopher Walters, and Edward Miguel. David Card, Jishnu Das, Stefano DellaVigna, Federico 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, and Stanford provided helpful feedback. We gratefully acknowledge generous funding and support by FCDO'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-0003835.

This is one of a series of working papers from "RISE"—the large-scale education systems research programme supported by funding from the United Kingdom's Foreign, Commonwealth and Development Office (FCDO), the Australian Government's Department of Foreign Affairs and Trade (DFAT), and the Bill and Melinda Gates Foundation. The Programme is managed and implemented through a partnership between Oxford Policy Management and the Blavatnik School of Government at the University of Oxford.

Please cite this paper as:

Andrabi, T. and Brown, C. 2022. Subjective versus Objective Incentives and Teacher Productivity. RISE Working Paper Series. 22/092. https://doi.org/10.35489/BSG-RISE-WP_2022/092

Use and dissemination of this working paper is encouraged; however, reproduced copies may not be used for commercial purposes. Further usage is permitted under the terms of the Creative Commons License.

The findings, interpretations, and conclusions expressed in RISE Working Papers are entirely those of the author(s) and do not necessarily represent those of the RISE Programme, our funders, or the authors' respective organisations. Copyright for RISE Working Papers remains with the author(s).

1 Introduction

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

In this paper, we ask two questions: What is the effect of subjective versus objective incentives on teacher productivity? Are subjective incentives able to help alleviate problems of noise and distortion, which often plague objective incentives? We answer these questions by conducting an 18-month randomized controlled trial with 234 private schools in Pakistan. We randomize schools to provide core teachers with one of three contracts: (i). control: flat raise – all teachers receive a raise of 5% irrespective of performance, (ii). treatment 1: subjective performance raise – teachers receive a raise from 0-10% based on their manager’s rating of their performance¹, or (iii). treatment 2: objective performance raise – teachers receive a raise from 0-10% based on their students’ mid-year and end of year test performance (Barlevy and Neal, 2012). Both treatments are within-school tournaments and have the same distribution of raise thresholds. These similarities allow us to isolate the effort response from just changing the performance metric (manager rating versus test score) while holding other features of the incentive structure constant.

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

¹Managers are generally principals or vice-principals and spend about a third of their time on employee management tasks, such as observations, feedback, and professional development.

²Items are drawn from the National Student Survey, Learning and Study Strategies Inventory, Big Five (children’s

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

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

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

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

We then draw on the model of moral hazard with multi-tasking to explain our main reduced form results: i). similar, positive effects of subjective and objective incentives on test scores, ii). negative effects of objective incentives on socio-emotional development, iii). significant differences in teacher

scale), Eisenberg’s Child-Report Sympathy Scale, Bryant’s Index of Empathy Measurement, Afrobarometer, World Values Survey, and Epistemic Curiosity Questionnaire.

classroom behavior across the two treatments. Moral hazard models with multi-tasking (Baker, 2002) isolate two main components of the incentive structure which affect employee response: noise (correlation between employee action and incentive pay) and distortion (correlation between piece rate for different actions and marginal return to those actions on firm outcomes). Our paper seeks to understand whether noise and distortion serve as important mechanisms of the reduced form effects we see.

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

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

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

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

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

To determine the effect of noise on student outcomes, we compare subjective treatment schools with managers whom teachers rate as accurate in assessing teacher effort versus managers rated as inaccurate in assessing teacher effort. We use this rating of managers’ accuracy interacted with

treatment status as an instrument for the perceived noisiness of the contract. We show that this rating of managers only affects teacher’s rating of noisiness in the subjective arm. This instrument for noise is robust to controlling for many other features of the contract and school environments.

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

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

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

Combined, these results help us understand why it is possible to have the same effect on test scores without needing to incentivize test scores directly. Subjective incentives are less noisy, producing a larger overall response, and less distorted, allowing teachers to prioritize multiple areas of student development. We find that the noise and distortion channel are able to explain a substantial portion of the reduced form effects we see.

Our paper makes three key contributions. First, it is the first study, to our knowledge, to isolate the causal effect of subjective versus objective incentives and the effect of subjective versus flat incentives for employees in any sector (Lazear and Oyer, 2012; Oyer and Schaefer, 2011). Existing studies have tested bundled incentives (a combined subjective and objective incentives versus no incentives) on employee behavior (Khan et al., 2019; Fryer, 2013). Previous work has also compared the effect of heterogeneity across plants to measure the effect of more or less steep

subjective incentives on employee overtime (Engelland and Riphahn, 2011). There is also evidence that managers, especially in educational settings, may have imperfect information about worker effort or may be biased toward certain groups (Jacob and Lefgren, 2008; Gibbs et al., 2004).

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

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

The remaining sections are organized as follows. Section 2 details the treatment and control conditions, the data collected, and standard implementation checks. Section 3 provides the main results of subjective and objective performance incentives on teacher effort and student outcomes. Section 4 gives an overview of the standard moral hazard model with multi-tasking and highlights the two key mechanisms which underpin the reduced form effects we find. Section 5 unpacks the mechanisms underlying the main effects in light of the moral hazard model, and section 6 concludes.

2 Experimental Design

2.1 Performance Incentive Treatments

We partnered with a large private school system in Pakistan to implement the research design. Schools are randomized to receive one of three contracts which determine the size of teachers’ raises at the end of the calendar year.³ 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

³Pairwise randomization by baseline test performance was used, which generally performs better than stratification for smaller samples (Bruhn and McKenzie, 2009).

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

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

- ***Subjective Treatment Arm:*** Teachers are evaluated by their manager at the end of the calendar year. Managers had complete discretion over how they evaluated teachers and what aspects of performance they would prioritize. To ensure teachers knew what was expected of them, managers delineated between 4-10 evaluation criteria, which would be used to evaluate the teachers. These included items such as improving their behavioral management of students, assisting with administrative tasks, helping plan an afterschool event, and improving students’ spoken English proficiency.⁴
- ***Objective Treatment Arm:*** Teachers are evaluated based on their average percentile value-added (Barlevy and Neal, 2012) for the spring and fall term. Percentile value-added is constructed by calculating students’ baseline percentile within the entire school system and then ranking their endline score relative to all other students who were in the same baseline percentile.⁵ We then average across all students the teacher taught during the two terms.

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

Both the subjective and objective treatment arms have several features in common, allowing us to isolate the effect of differing the performance metric and nothing else about the incentive structure. Both treatments are within-school tournaments, so this holds the level of competition fixed between the two treatments. In addition, the variance in the distribution of the incentive pay is equivalent across the two treatments. As we showed in section 4, holding the variance constant allows us to interpret differences in noise levels between the two systems as equivalent to differences in incentive

⁴An example set of criteria are provided in Appendix Table A1.

⁵Percentile 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.

steepness. The performance evaluation timeline also played out the same for all groups. Before the start of the year, managers set performance goals for their teachers irrespective of treatment. Teachers were evaluated based on their performance in January through December, with testing conducted in June and January to capture student learning in each term of the year.⁶

To ensure teachers and managers had full understanding of how each contract would work, we conducted an intensive information campaign with schools. First, the research team had an in-person meeting with each manager, explaining the contract assigned to their school, and, in the case of the subjective treatment, explaining what would be expected of them and when. Second, the school system’s HR department conducted in-person presentations once a term at each school to explain the contract. Third, teachers received frequent email contact from school system staff reminding them about the contract and half-way through the year contract teachers were provided midterm information about their rank based on the first 6 months.⁷ 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.

2.2 Timeline and Data

Our study was conducted from October 2017 through June 2019. It covered one performance review cycle conducted from January-December 2018 in which the contracts were in place. Figure 1 presents the main treatment implementation (detailed in section 2.1) and data collection activities (detailed below).

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

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

⁶The school systems’ central office designed and administered the June test to all students in a given grade. However, tests are graded locally by the school, often by the students’ teacher. Due to concerns of grade manipulation, grading was audited by the research team. 10% of all teacher’s exams were regraded. If the teachers’ grade and the auditor’s grade were off by more than 5%, another 10% of their tests were audited. If the average was still off by more than 5%, all of the teacher’s exams were regraded. Overall, grade manipulation was small and was generally driven by cases where teachers bumped up students’ grades from just failing to just passing. There was no heterogeneity in grading accuracy by treatment. The January test was conducted exclusively by the research team (described in section 2.2 below). These tests are *not* used as an outcome measure in this paper.

⁷An example midterm information note is provided in Appendix Figure A2.

classroom observations conducted by managers. Managers use a similar rubric to the one used by the research team to conduct classroom observations (detailed below).

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

Endline Survey: The teacher endline survey measured their understanding of the contract they were assigned, time use, and beliefs about their manager’s level of bias in conducting performance evaluations. The manager endline survey measured managers’ beliefs about teacher quality and measured management quality using the World Management Survey school questionnaire.⁸ The endline survey was conducted online with teachers and managers in spring and summer 2019. 6,080 teachers and 189 managers were surveyed.

Endline Student Testing and Survey: An endline test was conducted with students to measure performance in core subjects and socio-emotional skills after one year of the intervention. The research team conducted the endline test and student survey in January 2019. The test was conducted in Reading (English and Urdu), Math, Science, and Economics. The items were written in partnership with the school system’s curriculum and testing department to ensure appropriateness of question items. Grading was conducted by the research team. Items from international standardized tests (PISA, TIMSS, PERL, and LEAPS) and a locally used standardized test (LEAPS) were also included to benchmark student performance.⁹

Students also completed a survey to measure four areas of socio-emotional development. The areas are (i). love of learning (items drawn from National Student Survey, Learning and Study Strategies Inventory), (ii). ethical (items from Eisenberg’s Child-Report Sympathy Scale, Bryant’s Index of Empathy Measurement), (iii.) global citizen (items from Afrobarometer; World Values Survey), and (iv.) inquisitive (items from Learning and Study Strategies Inventory; Epistemic Curiosity Questionnaire). Appendix table A2 lists the survey items used for each area along with their source.

The choice of these four areas came from the school system’s priorities. They are the four areas of socio-emotional development they expect their teachers to focus on. These areas are posted on the walls in schools, and teachers receive professional development on these areas. Some managers also

⁸Due to budget constraints, we were unable to have the World Management Survey surveyors conduct the survey. Instead, we asked managers to directly rate themselves on the rubric that surveyors use. This approach could result in inflated management scores. As a result, we use additional objective data to corroborate the management scores.

⁹The endline student test data was used both for evaluating the effect of the treatments and used to compute objective treatment teachers’ raises.

specifically make these areas part of teachers’ evaluation criteria. In addition to these four areas, the survey also asked whether students liked their school or wanted to change to a different school.

Classroom Observation Data: To measure teacher behavior in the classroom, we recorded 6,800 hours of classroom footage and reviewed it using the Classroom Assessment Scoring System, CLASS (Pianta et al., 2012), which measures teacher pedagogy across a dozen dimensions.¹⁰ ¹¹ We also recorded whether teachers conducted any sort of test preparation activity and the language fluency of teachers and students.

Performance Evaluation Data: The school system had an existing performance evaluation system in which managers rated their teachers in December on performance criteria set in the previous December. We layered these new contracts on top of that existing system. In December 2017, before the announcement of treatments, managers set a number of performance criteria for each teacher, as they do each year. In a randomly chosen 3/4 of the subjective schools, those goals then become the evaluation criteria used to determine teachers’ raises for the following year. In the rest of the schools (objective, control, and the remaining subjective) those goals are used to provide feedback to teachers but have no financial consequence. In the remaining 1/4 of subjective schools, managers were required to create a new set of goals now that they knew there would be financial stakes attached to those goals. They were encouraged to set the goals to be focused on employee effort, rather than employee characteristics, like training or credentials. Since the performance evaluation system exists for all employees, we can use data on what goals were set and the scores on those goals to understand manager priorities and ratings with and without financial stakes tied to the performance rating.

2.3 Sample and Characteristics of the Employee-Manager Relationship

Teachers The study was conducted with a large private school system in Pakistan. The student body is from an upper middle-class and upper-class background. School fees are \$2,300-\$4,300 USD (PPP) per year. Teachers are generally younger and less experienced than their counterparts in the US, though they have similar levels of education. Table 1 presents summary statistics of our sample

¹⁰There are tradeoffs between conducting in-person observations versus recording the classroom and reviewing the footage. Videotaping was chosen based on pilot data which showed that video-taping was less intrusive than human observation (and hence preferred by teachers). Videotaping was also significantly less expensive and allowed for ongoing measurement of inter-rater reliability (IRR).

¹¹We did not hire the Teachstone staff to conduct official CLASS observations as it was cost-prohibitive and we required video reviewers to have Urdu fluency. Instead we used the CLASS training manual and videos to conduct an intensive training with a set of local post-graduate enumerators. The training was conducted over three weeks by Christina Brown and a member of the CERP staff. Before enumerators could begin reviewing data, they were required to achieve an IRR of 0.7 with the practice data. 10% of videos were also double reviewed to ensure a high level of ICC throughout the review process. We have a high degree of confidence in the internal reliability of the classroom observation data, but because this was not conducted by the Teachstone staff, we caution against comparing these CLASS scores to CLASS data from other studies.

compared to a representative sample of teachers in US (National Center for Education Statistics, 2011). Our sample is mostly female (80%), young (35 years on average), and inexperienced (5 years on average, but a quarter of teachers are in their first year teaching). All teachers have a BA and 68% have some post-BA credential or degree. Salaries are on average \$17,000 USD (PPP).

Managers In order to understand the effects of subjective performance pay, we need to understand who the managers are and what role they play in overseeing teachers. Managers here are either a principal in small schools or a vice principal in larger schools. They are tasked with overseeing the overall operations of the school and managing employees, including teachers and other support staff. Table 2 presents information about managerial duties compared to a US sample of principals. Like in the US, our managers are generally older (45 years old), less likely to be female (61%), and more experienced (9.6 years) than teachers. Most were previously teachers and transitioned into an administrative role. Managers spend about a 1/3 of their working hours overseeing their staff – observing classes, providing feedback, meeting with teachers and reviewing lesson plans. The rest of their time is spent on other tasks related to the schools functioning. The distribution of time use is fairly similar to the principals in the US.

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

2.4 Intervention Fidelity

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

Schools in the two treatment arms and control appear to be balanced along baseline covariates. Appendix Table A1 compares schools along numerous student and teacher baseline characteristics. Of 27 tests, one is statistically significant at the 10% level and one is statistically significant at the 5% level, no more than we would expect by random chance. Results presented include specifications which control for these few unbalanced variables.

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

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

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

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

3 Results

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

3.1 Effect of Incentives on Student Outcomes

3.1.1 Specification

Our main specification is:

$$Y_{i1} = \alpha + \beta_1 \textit{SubjectiveTreatment}_s + \beta_2 \textit{ObjectiveTreatment}_s + \delta Y_{i0} + \chi_j + \epsilon_i \quad (1)$$

The main dependent variable of interest is student outcome, Y_{i1} , for child, i , at endline, $t=1$. Student outcomes include test scores in Math, Science, English and Urdu and socio-emotional development. $\textit{SubjectiveTreatment}_s$ and $\textit{ObjectiveTreatment}_s$ are a dummy for whether the student's school, s , was assigned to subjective or objective performance raises. The left out group is the control group (flat raise). The coefficients of interest are β_1 and β_2 , and their test of equality. For test scores, we control for student's baseline score, Y_{i0} , to improve efficiency as there is high auto-correlation in test scores.¹² We also control for strata fixed effects, subject and grade, χ_j . Standard errors are clustered at the school level (the unit of randomization), and both standard and randomization inference p-values are provided in each table.

¹²For grade 4 students we do not have a baseline because standardized testing in the school system begins in 4th grade. For these students, and any other that are missing a baseline, we denote a score of zero and add a dummy for having the baseline test missing.

3.1.2 Results

Test Scores We find that both subjective and objective performance incentives have similar effects on test scores, of about 0.09 sd. Figure 2 and table 3 presents the results of each performance incentive on endline test scores. Column (1) shows results for all tests and question items. Effects are similar between the subjective and objective incentives, with an effect of 0.086 sd and 0.092 sd, respectively. In the row titled “F-test p-value (subj=obj)”, we present a test for the equality of $\beta_1 = \beta_2$. We cannot reject equality of effects between the two treatments on test scores. All results appear unchanged whether we consider standard p-values (in parentheses) or randomization inference p-values (in brackets).

Column (2) and (3) provide tests on the effect of the treatment by question item type to understand whether these effects are due to memorization of class content or actual learning. Column (2) only includes questions from the prior grade’s content and column (3) only includes questions that were added by the researchers from external standardized test sources including PISA, TIMSS, PERL and LEAPS.¹³¹⁴ Both sets of questions provide a useful test because it would not be possible for students to have memorized the answers to the questions. Remedial content (from previous grade levels) and external content are never tested on the school system’s standardized exam, and so teachers would not have prepared specifically for this material. Given that we find similar if not larger effects on these types of questions, it appears that treatment effects are coming from actual learning as opposed to memorizing curriculum. Again, we do not see a significant difference between the subjective or objective treatment.

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

Socio-Emotional Development While the effects on test scores were similar between both treatments, the effects on socio-emotional development paint a different picture. Figure 3 and table 4 presents the results on socio-emotional development overall and broken down socio-emotional area. Pooling across all five components of the index, we find the subjective treatment results in 0.05sd higher average score compared to the objective incentives, significant at the 10% level.¹⁵ When we split these results into their sub-areas, we see that the overall negative effect of objective incentives

¹³Not all subject and grade exams had remedial questions or external, so this is reflected in the decrease in sample size.

¹⁴Question items derived from these international sources were relevant to the curriculum of this school system and were not always matched to corresponding grade from the international exam, if that content was not part of the given year’s curriculum.

¹⁵These results come from a small negative effect of objective incentives relative to the control and a small positive effect of the subjective treatment relative to the control (neither of which are statistically significant).

relative to subjective is coming from a differential effect on “love of learning” and whether students like their school or would like to change schools. We can reject equality of the two treatments on these sub-areas at the 10% and 1% levels, respectively. This suggests that while objective incentives led to an increase in test scores, it was at the cost of enjoying school. Whereas, subjective incentives were able to accomplish the same learning gains without these negative consequences. On three other areas, ethical behavior, being a global citizen and inquisitiveness, we cannot reject the equality of the two treatments.

3.2 Effect of Incentives on Employee Effort

3.2.1 Specification

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

$$Y_i = \alpha + \beta_1 \textit{SubjectiveTreatment}_s + \beta_2 \textit{ObjectiveTreatment}_s + \chi_j + \epsilon_i \quad (2)$$

The main dependent variable of interest is outcome, Y_i , for teacher, i . Teacher outcomes include classroom observation scores and time use. We again control for grade and strata fixed effects, χ_j , and standard errors are clustered at the school level (the unit of randomization).¹⁶

3.2.2 Results

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

Figure 4 and table 5 presents the effects of each incentive on teachers’ overall classroom observation score, using the CLASS rubric. On average, objective teachers exhibit worse teaching pedagogy. They score 0.07pts lower on the 7pt CLASS rubric scale. Subjective teachers have no noticeable change in pedagogy quality, and we can reject the equality of the two treatments at the 10% level.

We then break down the 12 CLASS dimensions of pedagogy into three main areas, “class climate”, “differentiation”, and how “student-centered” the lesson is. “Class climate” captures whether the atmosphere of the classroom is positive, supportive and joyful or negative, punitive and dull. “Differentiation” captures whether the lesson is structured in a way to meet students who are

¹⁶We do not control for subject here, unlike in our student specification, because most teachers teach several subjects. In addition, for classroom observations, the observation period often overlapped with several subjects.

different proficiency levels and/or have different learning styles. Finally, “student-centered” measures how much of the lesson is teacher-directed versus student-involved. Teachers under the objective incentive contract have a more negative class climate and less student-centered lessons. Both see a decrease of around 0.1 pts. We can reject equality of treatments at the 10% level. There is also an increase in level of differentiation in the subjective and objective treatment schools.

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

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

One concern with classroom observation data is that teachers may worry the videos of their classrooms will be provided to their manager, and for subjective teachers that has more a consequence than for the other treatment arms. We do several things to help alleviate these concerns. First, in the consent form and during the camera set up, we communicate to teachers that the videos are confidential and will only be reviewed by the research team. We also let them know that only aggregated data at the school level will be provided to the school system head office. Second, visits were a surprise within a two-month window, so teachers could not adjust their lessons beforehand. Third, we recorded several hours back to back for each teacher. We find teachers are most aware (and responsive) of the camera in the first hour of taping. We can remove that data and repeat the same analysis and find very similar results.

Attendance and Time at Work We find that the subjective treatment results in a significant increase in the number of days a teacher is present at work relative to no incentives. Table 6 presents the results of the biometric clock in/out data. Relative to a control group mean of 145 days, subjective teachers are present an additional 6 days. We do not find an effect on hours spent at work for either treatment relative to the control. We cannot reject equality of treatments in either outcome. Columns (2) and (4) restrict to a sample of teachers who were present in the school system both terms and did not take any long leaves (health, maternity, etc.) to ensure the days present

result is not driven by these effects. Results are robust to this sample restriction.

3.3 How do Managers Implement the Subjective Incentive?

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

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

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

Our first piece of evidence on what managers value in teachers, comes from endline survey data from both teachers and managers. We asked both teachers and managers to respond to a hypothetical situation, in which a teacher asks them for advice about how to achieve a higher raise in the following year. They are then asked to rate how much time the teacher should spend on different types of actions. Table A3 presents the data from the survey question. Column 2 shows teachers’ responses about which actions would be most highly valued under the subjective contract. Column 3 presents responses to the same question posed to managers. Both subjective teacher and managers agree that improved pedagogy, like making lessons student centered and tailoring lessons to students at different initial levels, would increase their subjective rating. However, managers put additional weight on spending time collaborating with other teachers. Neither subjective teachers nor principals believe more superficial administrative tasks like volunteering at afterschool events or meeting with parents are important drivers of the subjective performance rating.

Our second piece of evidence also comes from the endline survey. We provide a vignette describing a hypothetical teacher to managers, and we ask them to provide a performance rating of the hypothetical teacher. The vignette randomizes the hypothetical teacher’s name, and rank in terms of value added, classroom behavioral management and attendance.¹⁷ Table A4 presents managers’

¹⁷The vignettes stated, “[Female name/Male name] is in the [bottom/middle/top] 10% of teachers in terms of

responses to this survey question. We find that managers highly value all three performance characteristics, but place double the weight on teacher value-added as they do on behavioral management and attendance. On average, moving from the 50th percentile value added to the 90th percentile value added would increase a teacher’s subjective rating by 0.7sd. Columns 1 through 3 of the table test each attribute separately. Columns 5 and 6 add all attributes together, and we see no difference in relative preference for these teacher characteristics. These results are also robust to adding manager fixed effects.

Finally, we can look at what teacher behaviors are correlated with teachers’ actual performance rating in the subjective treatment arm. Table A5 shows the relationship between teachers’ performance rating and teacher behaviors, as measured from classroom observation data, teacher value-added and biometric clock in/out data. We find that managers value higher value added and teacher attendance.¹⁸ This relationship remains when we control for subject and grade (column 2) and classroom observation scores (column 3). We find mixed evidence on the relationship between pedagogy and subjective rating. Some aspects of good pedagogy are valued (teachers who have a negative class climate have a lower rating) but others are not (teachers who spend more time on analysis/inquiry skills and have more student vs teacher talk time are negatively rated). One important limitation with this approach is that there are certainly omitted variables which we are unable to capture. However, having detailed classroom observation and time use data help us paint a relatively detailed picture of each teacher’s behavior. Combined these three pieces of evidence suggest that managers have preferences which are relatively aligned with the preferences of the school system.

Favoritism and bias A primary concern about subjective performance pay is whether managers are biased against certain employees or show favoritism toward preferred individuals. To assess whether this is a significant concern in this setting, we ask teachers at endline whether they felt their manager discriminated against certain groups or played favorites toward certain colleagues.¹⁹ Table A6 presents the results from these survey questions. On average, teachers in the subjective treatment arm are no more likely than teachers in the objective treatment arm to say that the contract unfairly favors certain teachers or that certain groups are discriminated against under this contract. Teachers also state that bias, gaming and favoritism is not a significant concern in either

students’ test score growth, in the [bottom/middle/top] 10% of teachers in terms of behavioral management, and is in the [bottom/middle/top]10% in terms of attendance and timeliness at work.” Managers rated three such vignettes with characteristics randomized across vignettes.

¹⁸There is a negative relationship between subjective rating and hours spent at school. This relationship may be driven by the fact that certain grades and teaching positions have different requirements about the length of the workday, so this could be picking up that variation rather than teacher effort.

¹⁹One concern with this approach is that teachers may be hesitant to provide honest assessment in a survey. To help minimize this concern teachers’ responses are anonymized and we communicate this to teachers at the time of consenting to the survey. We also ask the question several ways, including asking teachers to report such behavior about other schools or about the school system in general. This type of questions phrasing allows teachers to report problematic manager behavior while providing plausible deniability for their own manager.

contract.

Though teachers do not say that overt bias is a significant concern, we may be worried that there are more subtle types of bias at play. The primary type of bias we were concerned about in this setting is gender bias. In Pakistan, gender bias in employment is rampant (World Bank Group, 2018), and managers are more likely to be male than the employees they oversee. As part of the vignette survey questions, we include a way to test for subtle gender bias. In the vignettes we randomize the hypothetical teachers’ name to be a traditionally male or female Pakistani name. Table A4, column 3, presents the results of this test. We do not find that managers rate vignettes with female names lower.

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

Heterogeneity in treatment effects by manager characteristics On average the subjective treatment arm appears to have been successful at improving student outcomes and teacher effort, but there may be heterogeneity in how successfully managers implement the contract. We test for heterogeneity in treatment effects along several dimensions. First, table A7 presents heterogeneity in the subjective treatment arm by three manager characteristics: gender, age, and experience. We do not find significant differences in the effectiveness of the subjective treatment by these manager characteristics.

Second, table A7 presents heterogeneity in treatment effects by several dimensions of manager “quality”. We find that subjective performance pay is significantly less effective in schools where teachers believe their managers do not have an accurate perception of teacher effort. We measure this by asking teachers to rate how accurate their manager is in rating a fellow teacher.²⁰ We find there is no effect of subjective performance pay on student test scores for managers who are in the top quintile of this inaccuracy measure. We do not find heterogeneity and treatment effects by world management survey overall manager score (shown in Table A7, column (5)) or personnel management sub-score (column 6). However, as discussed in section 2.2, because this data was collected from manager self-report, we should be cautious about the interpretation, as managers may over rate themselves on these survey questions. This suggests that while subjective performance pay is on average very successful at producing learning games, these contracts may be ineffective in settings where employees do not trust their managers to implement them accurately.

²⁰To measure whether a manager has an accurate perception of what their teachers do, we ask teachers to answer the following question about three fellow teachers in their school, “The appraisal score their manager would give them is... [Too high/low by more than one raise category], [Too high/low by about one raise category], [Too high/low by less than one raise category], or [Accurate]”. We then construct an average of these ratings per manager;

4 Theoretical Framework

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

4.1 Moral Hazard with Multi-tasking

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

$$H(\mathbf{a}, e) = \mathbf{f} \cdot \mathbf{a} + e = f_1 a_1 + f_2 a_2 + \dots + e \quad (3)$$

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

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

$$P(\mathbf{a}, \phi) = \mathbf{g} \cdot \mathbf{a} + \phi = g_1 a_1 + g_2 a_2 + \dots + \phi \quad (4)$$

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

Teacher's utility is a function of their pay and a quadratic cost of effort.²¹

$$u(\mathbf{a}, \phi) = \mathbf{g} \cdot \mathbf{a} + \phi - \sum_{i=1}^n \frac{a_i^2}{2} \quad (5)$$

Teachers choose the optimal set of actions that maximizes their utility. Taking the derivative of Eq. 5, we have that the optimal decision is to set each action amount equal to the piece rate, $a_1^* = g_1, a_2^* = g_2, \dots, a_n^* = g_n$.

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

$$E[H(\mathbf{a}^*, e)] = \mathbf{f} \cdot \mathbf{g} = \|f\| \|g\| \cos \theta \quad (6)$$

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

We now go beyond Baker (2002) by making one additional assumption relevant in our context. We can further re-arrange the expression to show the effect that noise in the performance measure has on average human capital. Taking the variance of Eq. 4, we have $\text{var}(P) = \|g\|^2 \text{var}(\mathbf{a}) + \sigma_\phi^2$. Re-arranging, we can substitute this in for $\|g\|$ into Eq. 6. Average human capital then is:

$$E[H^*(\mathbf{a}^*, e)] = \|f\| \frac{\sqrt{\text{var}(P) - \sigma_\phi^2}}{\sqrt{\text{var}(\mathbf{a})}} \cos \theta \quad (7)$$

Here $\|f\|$ and $\text{var}(\vec{a})$ are constant across the two types of performance measures, subjective and objective, we will be comparing. In addition, due to the design of our subjective and objective incentives, $\text{var}(P)$, is also constant across the two schemes.²²

4.2 Theoretical Predictions

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

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

²¹Baker (2002) assumes risk-averse agents with a utility function of $u(\mathbf{a}, \phi) = E[P] - r\text{var}[P] - \sum_{i=1}^n \frac{a_i^2}{2}$. Because we are not focused on teacher retention, we leave out the risk aversion component, which only enters in determining the nature of the participation constraint and does not affect effort response once an employee has selected the contract.

²²A large class of incentives, including all tournaments, have a fixed variance, so the predictions of the model, apply in those cases as well.

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

Noise Noise captures how much of the performance incentive is unrelated to employee’s actions. This could operationalize as other factors outside the employee’s control affecting the performance measure (school resources, shocks, etc.) or mis-measurement of employee actions, if the contract attempts to measure teacher actions. It is important to flag that traditionally the way noise enters the optimal contract design is through reducing risk-averse employee’s utility. This requires firms to raise the fixed part of an employee’s salary to meet employee’s participation constraint. Here we are not focused on that consequence of noise as we are not focused on employee entry or exit in this paper²³.

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

4.3 Understanding the Experiment within the Theoretical Framework

An important distinction between our experimental performance pay system and this simple model is that our treatments are performance pay tournaments, where teachers are ranked relative to other teachers at their school and are allocated to one of five performance categories based on their relative performance. As a result, this framework is a simplification of the teacher’s problem in our experiment. However, the key predictions discussed above also hold in the more complicated tournament model, so we think it is illustrative of the mechanisms we will discuss in the next section.

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

²³A companion paper (Brown and Andrabi, 2020) studies employee sorting in response to these contracts

only certain aspects of human capital and not others, or most nefariously, they weight actions which make their job easier.

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

Our experiment connects to the model in two ways. First, in sections 5.3, we explicitly test the two predictions of the model using exogenous variation within one of the treatment arms that varies the level of noise and distortion. We then see the effect of these mechanisms on firm outcomes. Second, in section 5.2 and 5.4 we show that the difference in reduced form effects of each contract can be explained through differences in noise and distortion across the two contracts.

5 Mechanisms

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

First, in section 5.1, we present the similarities between the two treatments to help eliminate possible channels that could drive the difference in treatment effects. Second, in section 5.2, we highlight the differences between the systems. We show teachers believe subjective incentives to be less noisy and less distorted. Third, we provide evidence that noise and distortion does, in fact, affect outcomes. Section 5.3 shows that noise and distortion are related to student outcomes as predicted in the theoretical framework – more noise reduces the effect of incentives and more distortion diverts employee effort toward those actions. We conduct these tests by exploiting heterogeneity in levels of noise and distortion *within* a given treatment, to isolate the effect of noise or distortion on outcomes. Finally, in section 5.4, we bring together the estimates from section 5.2 and 5.3 to understand how much of the difference in the reduced form student effects can be explained by differences in noise and distortion.

5.1 Similarities between Treatments

In order to isolate the effect of the performance measure (percentile value-added versus manager rating), we hold a number of features constant between the two treatments. Both treatments are within-school tournaments. Both treatments provide a raise from 0-10% with the same set of rank

thresholds corresponding to raise amounts within that range. Both treatments were introduced at the same time in schools and had a similar performance review timing – manager completed midterm feedback in June 2018 and final ratings in December 2018 and the objective score was based on the average of tests in June 2018 and January 2019.

At endline, we survey teachers about their experience with their incentive scheme. We find no difference in teachers reported experience along a number of dimensions. There is no difference in their responses to the following survey questions: i). when teachers said they understood what was expected of them, ii). awareness of contract main features, iii). how frequently they thought about their contract, and iv). whether the system unfairly favors certain types of teachers (age, gender, etc). Figure 5 and table A6 provides results for each of these survey questions, showing no statistical difference between teachers’ responses by treatment.

5.2 Differences Across Treatments: Noise and Distortion

In this section through section 5.4, we will focus on two of the remaining differences between the treatments: noise and distortion. As highlighted in the theoretical framework, noise captures the extent to which a teacher’s actions affect their incentive payment. Distortion captures the extent to which actions which have the largest marginal return to human capital also are actions which have a higher effective piece rate under the given performance measure. First, we will show that the levels of noise and distortion are different across the treatments.

Noise We measure noise using teacher’s perceptions of the noisiness of their incentive treatment.²⁴ To measure perceived noise, we ask teachers to agree or disagree (on a 5pt scale), whether under their contract, “their raise is out of their control”, “those who work harder, earn more” and whether “I feel motivated to work harder”. Figure 6 presents the average response to each question with 1 being strongly disagree and 5 being strongly agree. We see that teachers in the subjective treatment, feel their raise is more in their control, hard work is rewarded, and they feel more motivated. The average difference is 0.14sd across the three areas, and we can reject equality of treatments for all three questions at the 5% level.

Distortion We measure distortion using endline survey data from teachers. We ask teachers to imagine a teacher who really wants to receive a higher raise at the end of the year and commits to work ten additional hours a week to increase their raise. Then we ask teachers how much of those ten hours should the teacher allocate different activities, such as collaborating with other teachers, incorporating higher order thinking skills into lessons, preparing practice tests, helping with extracurricular activities, etc. We then group these 17 different actions into four categories:

²⁴We think this is preferred to using “actual” noise, measured by seeing how predictive teacher’s measured behavior is to their raise. Perceived noise is what matters for teacher’s behaviors this last year, and there is likely measurement error that is correlated with treatment in measuring “actual” noise.

administrative tasks (grading, helping with extracurriculars, monitoring duty), professional development (collaboration, training, improved English skills and content knowledge), pedagogy (use of student-centered and differentiated lessons) and test preparation (achieving certain grade targets).

We find that teachers in subjective versus objective schools feel that there are some slight differences in which actions should be prioritized in order to increase their raise. Figure 7 and table 7 presents the differences in stated valuation of each area. Overall, teachers think those under the subjective contract should prioritize more administrative tasks and slightly less on test preparation. We will show in the next section that these actions have different implications for student outcomes.

5.3 Effect of Noise and Distortion on Outcomes

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

To test whether noise affects outcomes, we exploit heterogeneity *within* the subjective treatment in noisiness. Managers vary in their accuracy of assessing teacher effort. Some managers observe lessons for each of their teachers every week. Others sit down and review paper lesson plans, and some are more hands off. To measure whether a manager has an accurate perception of what their teachers do, we ask teachers to answer the following question about three fellow teachers in their school, “The appraisal score their manager would give them is... [Too high/low by more than one raise category], [Too high/low by about one raise category], [Too high/low by less than one raise category], or [Accurate]”. We then construct an average of these ratings per manager, capturing average *perceived* inaccuracy. On average, teachers believe their managers over or under rate their fellow teachers by 0.8 of an appraisal step (out of the five-step system shown in section 2.1. However, there is considerable heterogeneity. Those most inaccurate quintile of managers are perceived to rate other teachers incorrectly by greater than two steps.

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

We find that $ManagerRatingInaccuracy_j$ significantly predicts teacher’s rating of the noisiness of their appraisal system in subjective but not objective/control schools, as we would expect. A 1 sd increase in manager inaccuracy increases beliefs about the noisiness of the contract by 0.1-0.4 sd in subjective schools. Table 8 presents the results from the first stage for data at the teacher and student

level²⁵. Columns (2) and (4) add additional controls, including teacher’s beliefs about the preference for different actions (“distortion”) and teacher beliefs about other non-noise features of the contract (timing, understanding, etc). The coefficient on *ManagerAccuracy* * *SubjectiveTreatment* is very robust to the inclusion of these controls, suggesting that this instrument is picking up difference in noise and not other features of the contract environment.

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

$$\begin{aligned} Outcome_{ij} = & \alpha_0 + \alpha_1 ManagerRatingInaccuracy_j + \alpha_2 SubjectiveTreat_i \\ & + \alpha_3 \widehat{Noise} + \chi_{ij} + \epsilon_{ij} \end{aligned} \quad (8)$$

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

We find that noise significantly reduces the effectiveness of performance incentives (table 9). A 1 sd increase in noisiness of the incentive scheme reduces teachers’ hours worked by 13.2 hours per week and reduces test scores by 0.175 sd. We do not find an effect of noise on socio-emotional scores. Because our effective first stage has an f-stat of less than 10, we present the AR test p-values which are our preferred test, given that they are robust to weak instruments in the just-identified case. Columns (2), (4), and (6) add in the same additional controls as in table 8 for non-noise features of the contract environment. The effect of noise on hours worked and test scores is robust to the addition of those controls.

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

²⁵The first stage in table 8 columns (1) and (2) is at the level of the teacher used for the hours results in table 9, column (1) and (2). The first stage in columns (3) and (4) is at the level of the student and used for the student test and socio-emotional skills outcomes in table 9, column (3)-(6).

$$\begin{aligned}
StudentOutcome_i = & \alpha_0 + \alpha_1 SubjectiveTreat_i + \sum_{j=1}^J \delta_j Points\ on\ Action\ j_i \\
& + \sum_{j=1}^J \beta_j Points\ on\ Action\ j_i * SubjectiveTreat_i + \chi_{ij} + \epsilon_i
\end{aligned} \tag{9}$$

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

We find that several of the action categories are related to student outcomes. Table 10 presents the β_j 's for each action category. Professional development and test prep actions are positively related to student test scores. However, test prep is negatively related to student socio-emotional scores. These results are robust to the inclusion of additional controls about the contract environment (table 10, column (2) and (4)).

5.4 Contribution of Noise and Distortion to Reduced Form Effects

Finally, we can pull the results together to understand the extent to which noise and distortion can explain the reduced form results we saw in section 3.1. To do this we decompose the total reduced form effect into the component from noise, distortion and an unexplained component, ϵ :

$$\begin{aligned}
dTestScore = & \frac{\partial TestScore}{\partial Noise} * dNoise + \frac{\partial TestScore}{\partial Distortion} * dDistortion + \epsilon \\
-0.006sd = & -0.17 * -0.14sd + -0.03sd + \epsilon \\
\epsilon = & 0.0002sd
\end{aligned} \tag{10}$$

The overall effect of subjective relative to objective on test scores was close to zero (-0.006sd, from table 3). The effect of noise on test scores is -0.17 (table 9) and there is 0.14sd less noise in the subjective arm than the objective arm (figure 6). For the distortion component, we repeat the same approach for each of the four action categories (admin, professional development, pedagogy and test prep). We take the difference between subjective and objective for each area (table 7), multiply each category with the return to preference for that action on test scores (table 10) and sum. In total, $\frac{\partial TestScore}{\partial Distortion} * dDistortion$, then is -0.03. Subjective schools put slightly less focus on test scores. Combined, the positive effect of subjective having less noise and the negative effect of them placing

less focus on test scores almost cancel each other out. Overall, the remaining unexplained portion, ϵ , is just 0.0002sd, suggesting noise and distortion are effective at explaining the student results.

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

$$\begin{aligned}
dSEScore &= \frac{\partial SEScore}{\partial Noise} * dNoise + \frac{\partial SEScore}{\partial Distortion} * dDistortion + \epsilon \\
0.0433sd &= -0.06 * -0.14sd + 0.011sd + \epsilon \\
\epsilon &= 0.024sd
\end{aligned} \tag{11}$$

The overall effect of subjective relative to objective on socio-emotional development was 0.0433 sd (table 4). The effect of noise on socio-emotional skills is -0.06 and there is -0.14sd less noise in the subjective arm than the objective arm. The subjective teachers focus more on tasks which are related to socio-emotional skills. Overall $\frac{\partial TestScore}{\partial Distortion} * dDistortion$ is 0.011 sd. The remaining unexplained portion is 0.024 sd, or about half of the difference between the subjective and objective treatment. This is perhaps unsurprising given the results throughout this section. Noise and distortion were much less related to socio-emotional skills than test scores. This could be because there is in fact a weak relationship between them. Alternatively, we may not be as successful at measuring socio-emotional skills and certainly have a harder time capturing what aspects of teacher’s behavior is related to developing these skills. Better measurement along these areas is an important area for future work.

6 Conclusion

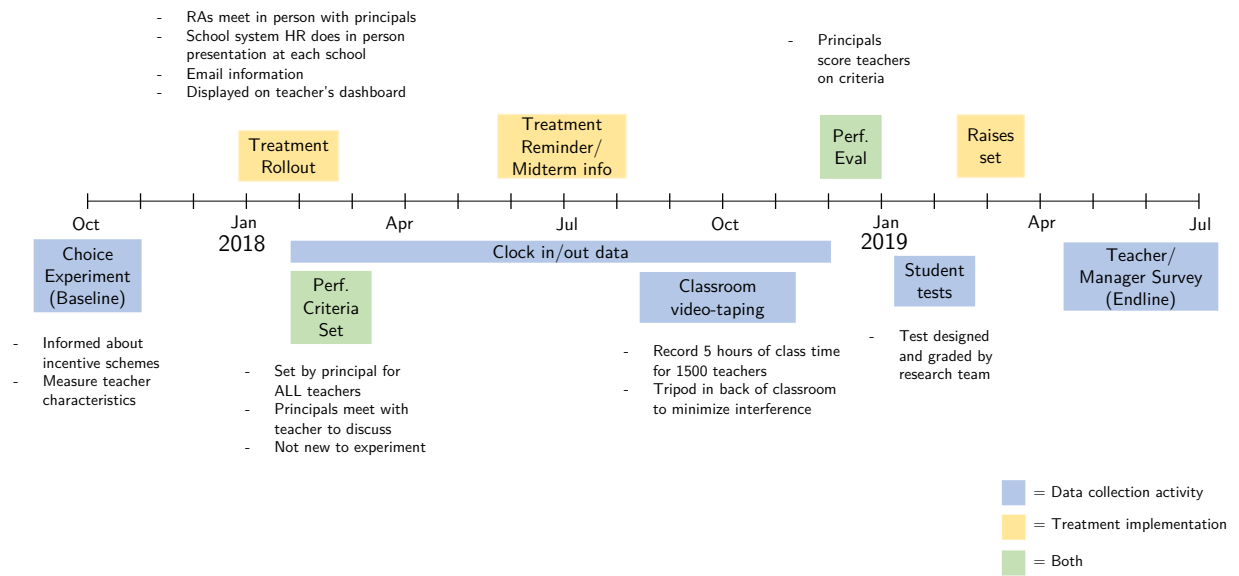
In this paper, we provide evidence on the effect of subjective versus objective incentives for teachers. We find that both subjective and objective incentives increase test scores, but objective incentives result in negative effects on socio-emotional development. These student outcomes make sense given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

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

but a smaller fraction of the reduced form socio-emotional skill effect.

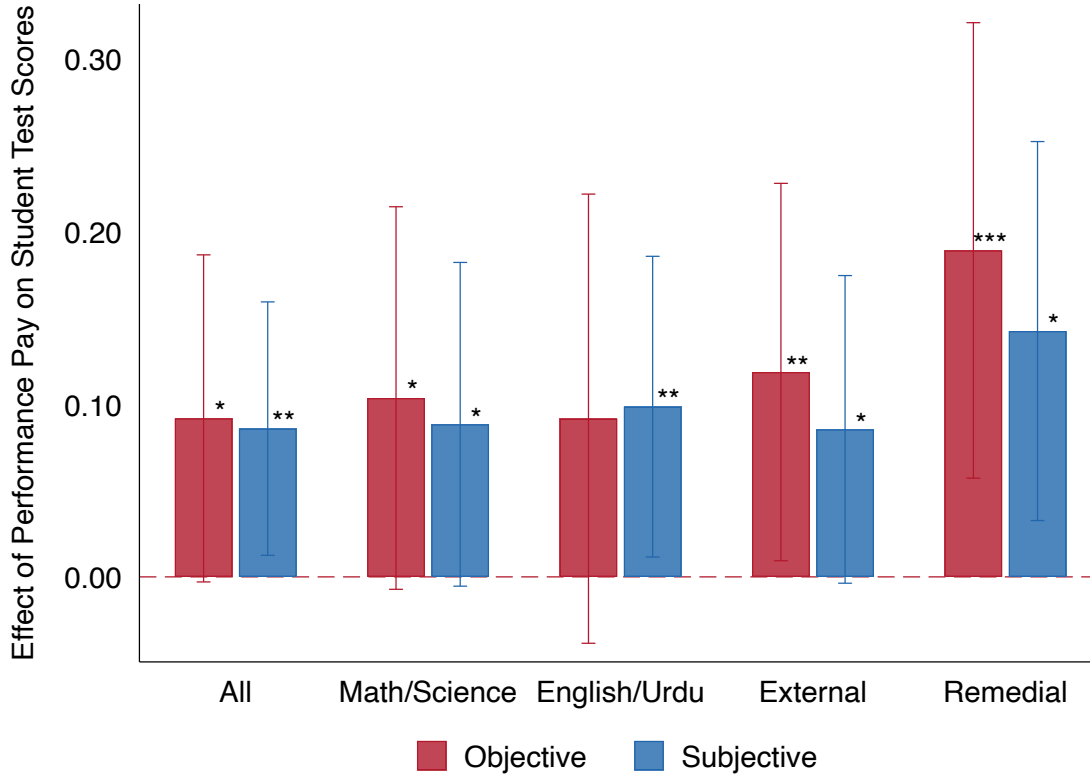
7 Figures

Figure 1: Experimental Timeline



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

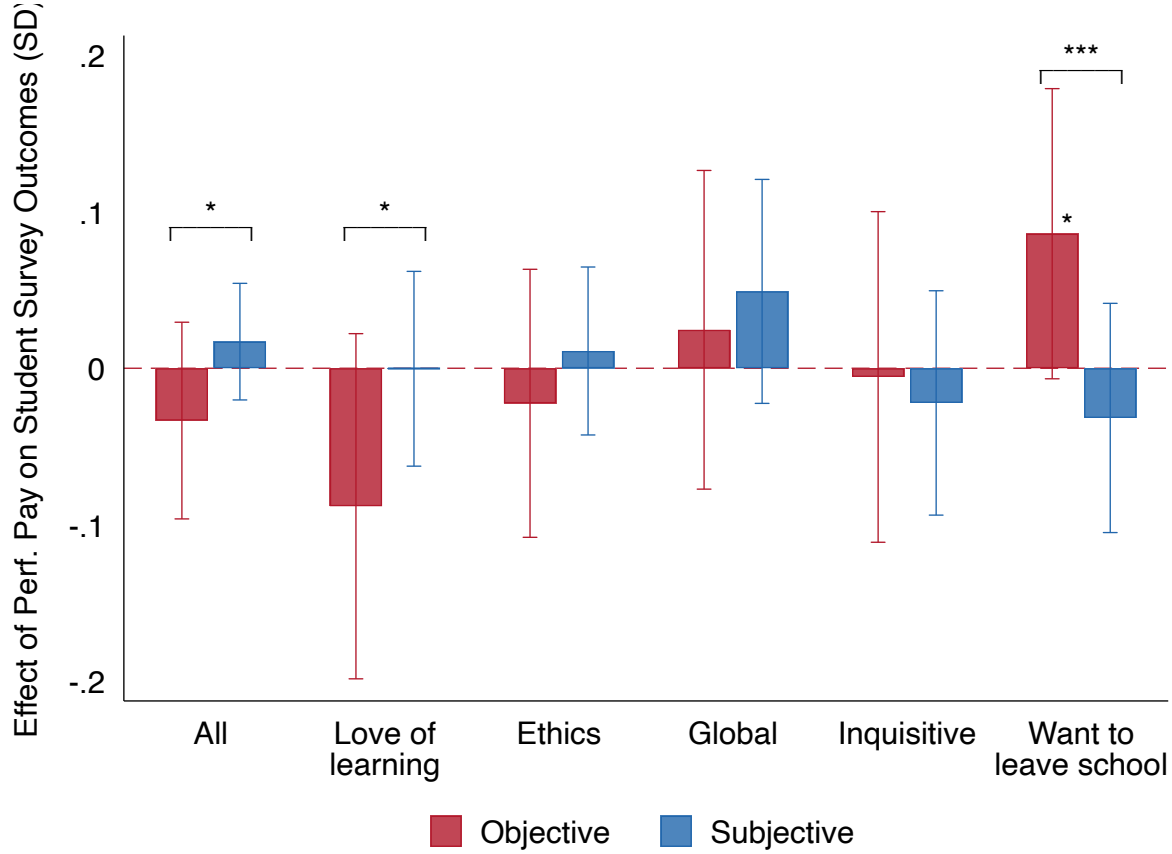
Figure 2: Effect of Incentives on Student Test Scores



Notes: This figure presents the effects of each performance incentive treatment on student endline test scores relative to the control group.

- The blue bars present the coefficient of the effect of the objective treatment relative to the control (flat raises). The red bars present the coefficient of the effect of the subjective treatment relative to the control.
- The observation is at the student-subject level. The y-axis presents the coefficient from a regression of student's z-score on a given endline exam on treatment dummy variables.
- The sample includes students tested in grades 4-13 in five subjects: Math, Science, English, Urdu, Economics.
- The first two bar graphs includes all test subjects and question items. The next two bars restricts to math and science exams. The next restrict to English, Urdu and Economics exams. The next restrict to question items drawn from external sources, such as PISA and TIMSS. The last two restrict to question items which were from the previous grade.
- All regressions include strata fixed effects and control for baseline student average test score, baseline school average test score, grade and subject. Standard errors are clustered at the school level. 95% confidence intervals are shown on each bar. Stars just above a bar show the significance of the treatment group relative to the control. A bracket above two bars denotes the significance between the two treatments (subjective versus objective). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

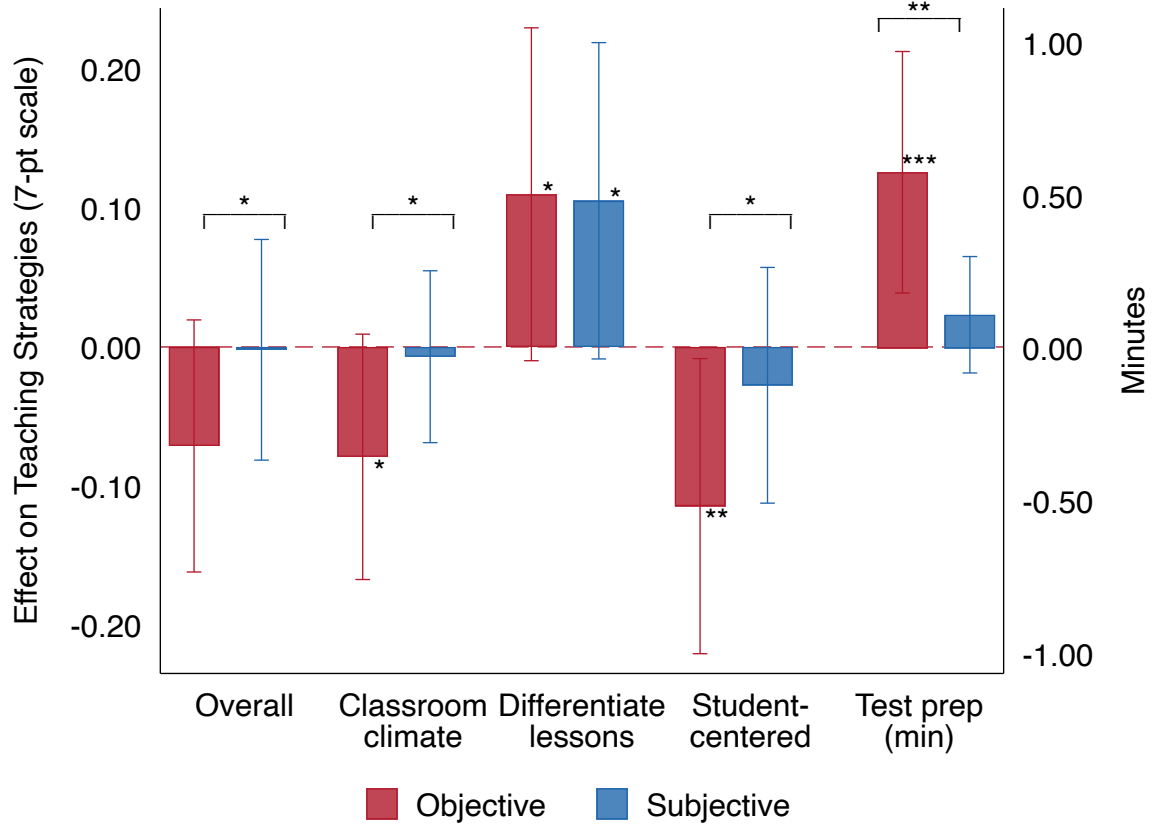
Figure 3: Effect of Incentives on Socio-Emotional Outcomes



Notes: This figure presents the effects of each performance incentive treatment on student socio-emotional outcomes relative to the control group.

- The blue bars present the coefficient of the effect of the objective treatment relative to the control (flat raises). The red bars present the coefficient of the effect of the subjective treatment relative to the control.
- The observation is at the student level. The y-axis presents the coefficient on a regression of student's z-score on a given socio-emotional dimension from an endline survey of students conducted in January 2019.
- The first two bars provides the average across all five dimensions of socio-emotional outcomes. The remaining ten provide effects on each individual dimension.
- All regressions include strata fixed effects and control for student's grade. Standard errors are clustered at the school level. 95% confidence intervals are shown on each bar. Stars just above a bar show the significance of the treatment group relative to the control. A bracket above two bars denotes the significance between the two treatments (subjective versus objective). $*p < 0.10$, $**p < 0.05$, $***p < 0.01$.

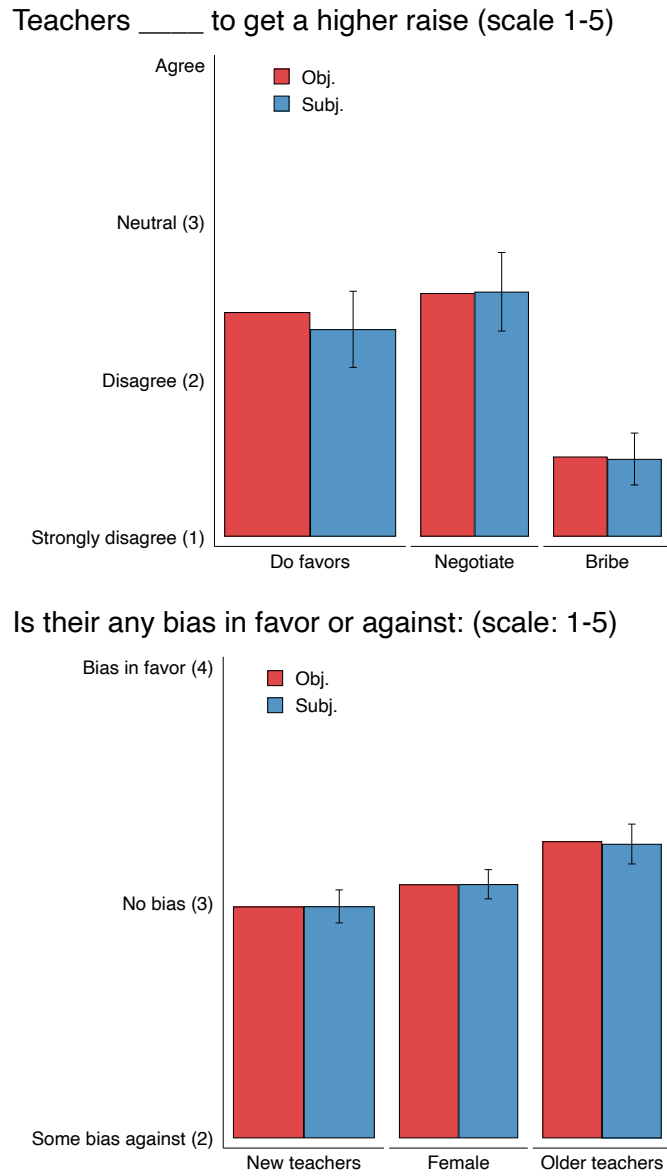
Figure 4: Effect of Incentives on Teacher Pedagogy



Notes: This figure presents the effects of each performance incentive treatment on teacher behavior as rated based on classroom videos relative to the control group.

- The blue bars present the coefficient of the effect of the objective treatment relative to the control (flat raises). The red bars present the coefficient of the effect of the subjective treatment relative to the control.
- The observation is at the classroom observation level. Teachers may be observed multiple times over the course of the intervention. The y-axis presents the coefficient from a regression of classroom observation score in a given dimension on treatment dummy variables.
- The sample includes teachers from grades 4-13 in core academic subjects.
- The first two bars presents the effects on the average score on the CLASS rubric (Pianta et al., 2012), on a 7-pt scale. The next six bars provide effects on scores on sub-areas of the class rubric. The last two bars provide effects on time spent on testing or test-prep activities (in minutes).
- All regressions include strata fixed effects and control for grade and video coder fixed effects. Standard errors are clustered at the school level. 95% confidence intervals are shown on each bar. Stars just above a bar show the significance of the treatment group relative to the control. A bracket above two bars denotes the significance between the two treatments (subjective versus objective). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

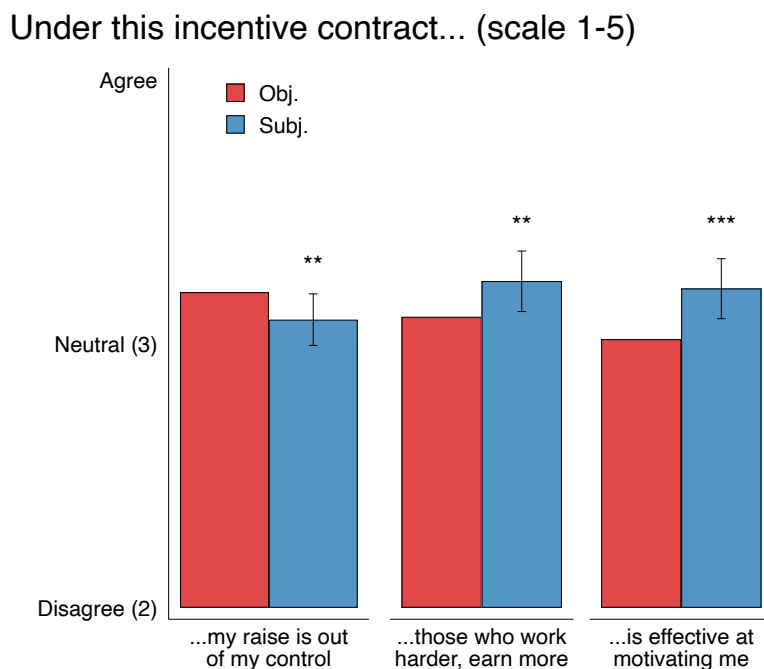
Figure 5: Beliefs about Contracts by Treatment



Notes: This figure presents teacher's responses to questions regarding their incentive contract for the previous year.

- Figure A shows teachers responses to questions about what actions they believe fellow teachers take to increase their raise. Figure B shows their responses to questions about whether certain groups are favored by the incentive scheme.
- The red (blue) bars presents the average response for teachers in the objective (subjective) treatment schools. The observation is at the teacher level and come from the endline survey of teachers.
- In figure A, the outcome is a 5pt scale from Strongly Disagree (1) to Strongly Agree (5). In figure B, the outcome is a 5pt scale (1, lots of bias against, 3, no bias, 5, lots of bias in favor).
- Standard errors are clustered at the school level. 95% confidence intervals are shown on the subjective bar comparing it to the objective treatment. Stars just above the bar show the significance of the subjective group relative to the objective. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 6: Difference in Noise by Treatment

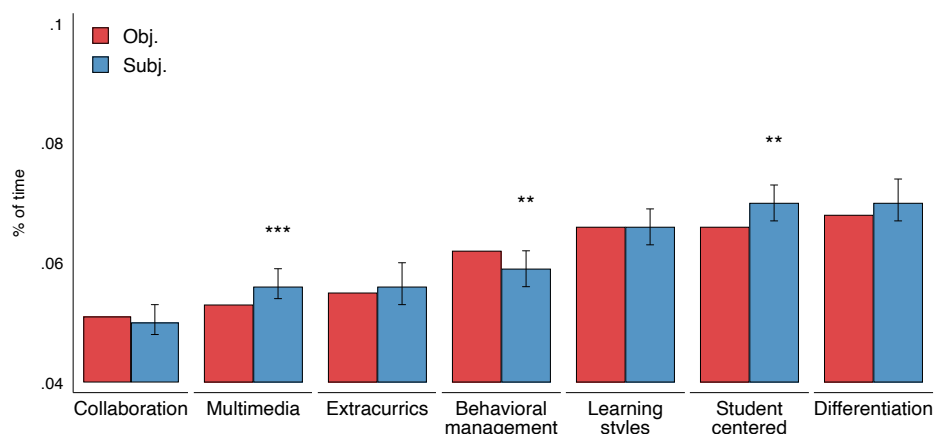


Notes: This figure presents teacher's responses to a questions about how they respond to their incentive.

- The red (blue) bars presents the average response for teachers in the objective (subjective) treatment schools. The observation is at the teacher level and come from the endline survey of teachers.
- The questions are on a 5pt scale from Strongly Disagree (1) to Strongly Agree (5).
- Standard errors are clustered at the school level. 95% confidence intervals are shown on the subjective bar comparing it to the objective treatment. Stars just above the bar show the significance of the subjective group relative to the objective. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 7: Difference in Value of Teacher Activity by Treatment

To increase my raise I should spend time on...



Notes: This figure presents teachers' responses to a hypothetical scenario in which they are advising a teacher which actions they should take to increase their raise under a given treatment.

- The red (blue) bars presents the average response for teachers in the objective (subjective) treatment schools. The observation is at the teacher level and come from the endline survey of teachers.
- The questions are on a 5pt scale from Strongly Disagree (1) to Strongly Agree (5).
- Standard errors are clustered at the school level. 95% confidence intervals are shown on the subjective bar comparing it to the objective treatment. Stars just above the bar show the significance of the subjective group relative to the objective. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

8 Tables

Table 1: Descriptive Statistics about Teachers in Study and Comparison Sample

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

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

Table 2: Descriptive Statistics about Mangers in Study and Comparison Sample

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

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

Table 3: Effect of Incentives on Student Test Scores

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

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

Table 4: Effect of Incentives on Student Socio-Emotional Outcomes

	Endline Survey Indices (z-score)					
	All (1)	Love of learning (2)	Ethical (3)	Global (4)	Inquisitive (5)	Dislike school (6)
Objective Treatment	-0.0334 (0.294) [0.395]	-0.0882 (0.116) [0.0970]	-0.0224 (0.605) [0.670]	0.0246 (0.634) [0.670]	-0.00556 (0.917) [0.930]	0.0860* (0.0688) [0.129]
Subjective Treatment	0.0170 (0.368) [0.560]	-0.000335 (0.992) [0.993]	0.0110 (0.685) [0.762]	0.0491 (0.177) [0.228]	-0.0222 (0.542) [0.629]	-0.0318 (0.392) [0.451]
F-test pval (subj=obj)	0.09	0.08	0.40	0.58	0.72	0.00
Randomiz infer pval (subj=obj)	0.0940	0.0250	0.511	0.609	0.726	0.00800
Control Group Mean	-0.00	-0.00	-0.00	0.00	-0.01	0.38
Clusters	126	126	126	125	126	124
Observations	15418	15401	14904	14168	14909	11505

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

Table 5: Effect of Incentives on Teacher Effort

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

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

Table 6: Effect of Teacher Time at Work

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

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

Table 7: Teachers Perceptions about which Actions to Focus on by Treatment

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

Notes: This table reports teachers' responses to a hypothetical scenario in which they are advising a teacher which actions they should take to increase their raise under a given treatment. Data was collected as part of the endline survey, and observations are at the unit of the teacher. Actions are categorized into four categories: administrative tasks, pedagogy, professional development, and test preparation. Table A3 provides teacher's weight for the full list of activities by treatment. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 8: Instrumenting Noise with Manager Accuracy - First Stage

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

Notes: This table presents the relationship between manager rating inaccuracy and teacher's rating of how noisy their contract was. The outcome is teacher's rating of how noisy their contract was as measured by an index of their response to the three questions shown in Figure 6. Columns (1) and (2) uses data at the teacher level. Columns (3) and (4) uses data at the teacher-student exam level. Student exam data is matched to all teachers who taught the student in the given exam subject for at least one term from January-December 2018. All regressions control for subject, class and manager inaccuracy squared. Columns (3) and (4) also control for school and student test baseline. Columns (2) and (4) add in additional controls to pick up other non-noise differences across contracts. These controls include weight placed on each of the four activity groups listed in Table 7, those values interacted with the Subjective treatment, when teachers said they learned about the treatment and how often they received information about the treatment. Standard errors are clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 9: Effect of Noise on Outcomes

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

Notes: This table presents the relationship between teacher's rating of the noisiness of their contract, instrumented by manager inaccuracy*Subjective Treatment, on teacher and student outcomes. Columns (1) and (2) use data at the teacher level. Columns (3) and (4) use data at the teacher-student exam level. Student exam data is matched to all teachers who taught the student in the given exam subject for at least one term from January-December 2018. Columns (5) and (6) uses data the student level. All regressions control for subject, class, subjective treatment, manager inaccuracy, and manager inaccuracy squared. Columns (3) and (4) also control for school and student test baseline. Columns (2), (4) and (6) add in additional controls to pick up other non-noise differences across contracts. These controls include weight placed on each of the four activity groups listed in Table 7, those values interacted with the Subjective treatment, when teachers said they learned about the treatment and how often they received information about the treatment. Standard errors are clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 10: Effect of Manager Preferences on Student Outcomes

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

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

9 References

References

- Baker, George**, “Distortion and Risk in Optimal Incentive Contracts,” *The Journal of Human Resources*, 2002, *37* (4), 728.
- Barlevy, Gadi and Derek Neal**, “Pay for Percentile,” *American Economic Review*, August 2012, *102* (5), 1805–1831.
- 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/eco.j.12267>.
- Brown, Christina and Tahir Andrabi**, “Overcoming Adverse Selection through Performance Pay,” *Working Paper*, July 2020.
- Bruhn, Miriam and David McKenzie**, “In Pursuit of Balance: Randomization in Practice in Development Field Experiments,” October 2009, *1* (4), 34.
- Chetty, Raj, John N Friedman, and Jonah E Rockoff**, “Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates,” *American Economic Review*, September 2014, *104* (9), 2593–2632.
- Engellandt, Axel and Regina Riphahn**, “Evidence on Incentive Effects of Subjective Performance Evaluations,” *Industrial & Labor Relations Review*, 2011, *64*.
- 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.
- Gibbs, Michael, Kenneth A. Merchant, Wim A. Van der Stede, and Mark E. Vargus**, “Determinants and Effects of Subjectivity in Incentives,” *The Accounting Review*, 2004, *79* (2), 409–436.
- 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.
- Jacob, Brian A. and Lars Lefgren**, “Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education,” *Journal of Labor Economics*, January 2008, *26* (1), 101–136.

- Khan, Adnan Q., Asim Ijaz Khwaja, and Benjamin A. Olken**, “Making Moves Matter: Experimental Evidence on Incentivizing Bureaucrats through Performance-Based Postings,” *American Economic Review*, January 2019, *109* (1), 237–270.
- Lavy, Victor**, “Using Performance-Based Pay to Improve the Quality of Teachers,” *The Future of Children*, 2007, *17* (1.), 87–109.
- Lazear, Edward P and Paul Oyer**, “Personnel Economics,” in “The Handbook of Organizational Economics,” Princeton; Oxford: Princeton University Press., December 2012, pp. pp. 479–519.
- Leigh, A.**, “The Economics and Politics of Teacher Merit Pay,” *CESifo Economic Studies*, March 2013, *59* (1), 1–33.
- Muralidharan, Karthik and Venkatesh Sundararaman**, “Teacher Performance Pay: Experimental Evidence from India,” *Journal of Political Economy*, February 2011, *119* (1), 39–77.
- National Center for Education Statistics**, *Schools and Staffing Survey, 2010-2011: [United States]*, U.S. Dept. of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 2011.
- Oyer, Paul and Scott Schaefer**, “Personnel Economics: Hiring and Incentives,” in O. Ashenfelter and D. Card, eds., *O. Ashenfelter and D. Card, eds.*, 1 ed., Vol. 4B, Elsevier, 2011, pp. 1769–1823.
- 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.
- Podgursky, Michael J. and Matthew G. Springer**, “Teacher performance pay: A review,” *Journal of Policy Analysis and Management*, 2007, *26* (4), 909–950.
- Prendergast, Canice**, “The Provision of Incentives in Firms,” *Journal of Economic Literature*, 1999, *37* (1), 7–63.
- , “The Motivation and Bias of Bureaucrats,” *The American Economic Review*, 2007, *97* (1), 18.
- **and Robert Topel**, “Discretion and bias in performance evaluation,” *European Economic Review*, April 1993, *37* (2-3), 355–365.
- World Bank Group**, *Country Policy And Institutional Assessment Dataset* 2018.

10 Appendix Figures

Figure A1: 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 A2: 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

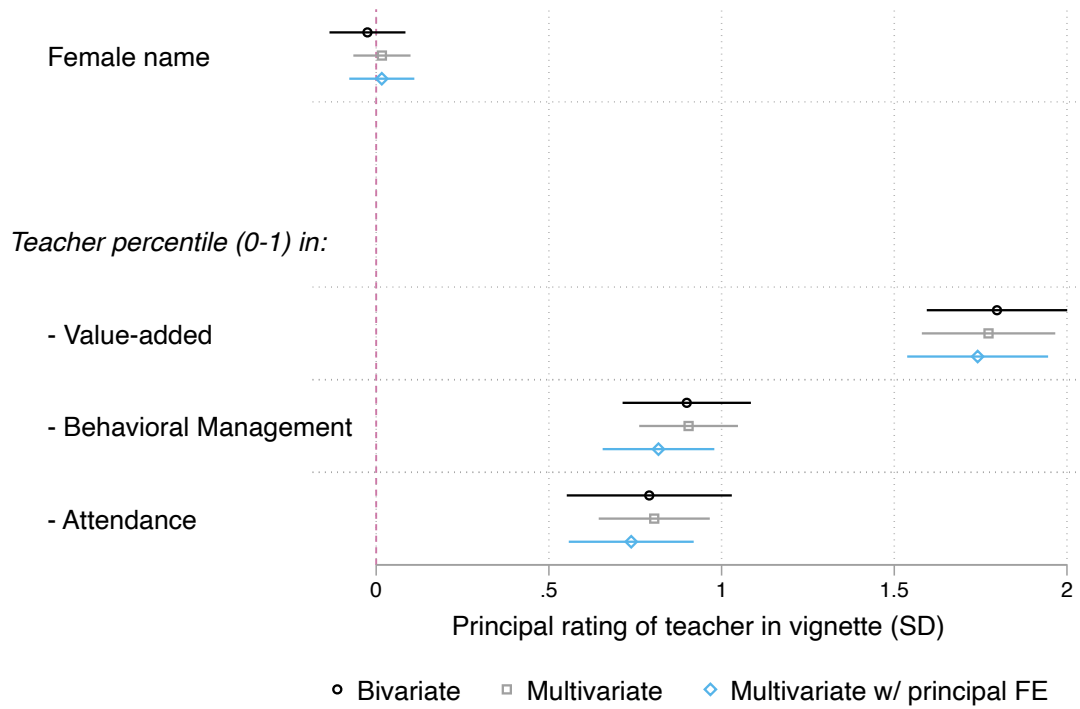
Your midterm performance is:

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.

Figure A3: Manager Rating by Vignette Characteristics



Notes: These figures presents the relationship between vignette characteristics and principal's rating of the teacher in the vignette. Principals received the vignettes stating, "[Female name/Male name] is in the [bottom/middle/top] 10% of teachers in terms of students' test score growth, in the [bottom/middle/top] 10% of teachers in terms of behavioral management, and is in the [bottom/middle/top]10% in terms of attendance and timeliness at work." Managers rated three such vignettes with characteristics randomized across vignettes. *Teacher Value-Added Percentile*, *Teacher Behavioral Management Percentile*, and *Teacher Behavioral Management Percentile* takes values, 10, 50 and 90 to correspond to the bottom, middle and top 10% listed in the vignette. *Principal rating of teacher in vignette (SD)* is the principal's stated rating of the teacher in the vignette in standard deviations.

11 Appendix Tables

Table A1: Baseline Covariates

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

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

Table A2: 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 A3: Percent of Time Individuals Believe Should be Spent on Each Type of Activity

Variable	(1) Objective Teachers N/ [Clusters] Mean/ SE	(2) Subjective Teachers N/ [Clusters] Mean/ SE	(3) Subjective Managers N/ [Clusters] Mean/ SE	(1)-(2)	T-test Difference (1)-(3)	(2)-(3)
Improving behavioral management	487 (0.001)	2406 (0.001)	41 (0.006)	0.003**	0.009*	0.006
Collaborating with other teachers	487 (0.001)	2406 (0.000)	41 (0.005)	0.001	-0.008*	-0.009**
Grading student papers	487 (0.002)	2406 (0.001)	41 (0.005)	-0.003	-0.002	0.001
Providing differentiated lessons	487 (0.002)	2406 (0.001)	41 (0.005)	-0.003	0.000	0.003
Helping with extracurriculars	487 (0.002)	2406 (0.001)	41 (0.005)	-0.001	0.008	0.009
Incorporating higher order thinking skills	487 (0.002)	2406 (0.001)	41 (0.005)	0.001	-0.000	-0.001
Catering to different learning styles	487 (0.001)	2406 (0.001)	41 (0.005)	0.000	0.001	0.001
Incorporating multimedia	487 (0.001)	2406 (0.001)	41 (0.006)	-0.004**	-0.000	0.003
Communicating with parents	487 (0.001)	2406 (0.001)	41 (0.004)	0.002	0.001	-0.002
Conducting practice tests	487 (0.002)	2406 (0.001)	41 (0.007)	0.002	-0.001	-0.003
Making lessons more student centered	487 (0.001)	2406 (0.001)	41 (0.007)	-0.003**	-0.017***	-0.013***

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

Table A4: Manager Rating by Vignette Teacher Characteristic

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

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

Table A5: Teacher Effort and Subjective Performance Rating

	Subjective Performance Rating Percentile (0-100)		
	(1)	(2)	(3)
Hours present at school	-1.793*** (0.293)	-1.550*** (0.306)	-1.979*** (0.617)
Days present at school	0.167*** (0.0263)	0.153*** (0.0263)	0.232*** (0.0628)
Value-Added	1.393** (0.575)	1.574*** (0.581)	3.417** (1.388)
CLASS Rubric Dimensions:			
Positive Climate			-7.472* (3.836)
Teacher Sensitivity			0.323 (3.647)
Regard for Student Perspectives			1.650 (1.847)
Behavioral Management			0.282 (3.574)
Productivity			-3.829 (3.492)
Negative Climate			-12.49* (6.865)
Instructional Learning Formats			5.060 (3.396)
Content Understanding			1.780 (3.051)
Analysis and Inquiry			-4.815** (2.268)
Quality of Feedback			3.681 (2.791)
Student Talk Time			-5.721** (2.427)
Student Engagement			5.804 (3.651)
Other aspects of classroom observation:			
Students Use of English			-0.0498 (0.0747)
Classroom is decorated			-0.659 (6.348)
Use of technology			-0.0803 (0.780)
Time spent on test prep			0.978 (1.310)
Observations	2778	2628	618
Dependent Variable Mean	49.05	49.05	49.05
Subject and Grade Controls		X	X

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

Table A6: Teacher's beliefs about contract features

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

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

Table A7: Heterogeneous Treatment Effects by Manager Characteristics

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

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