Understanding Gender Discrimination by Managers*

Christina Brown†

April 29, 2022

Abstract

Pakistan ranks in the lowest decile in female labor force participation, and even in sectors where women are more prevalent, such as teaching, they earn 70 cents for each dollar men earn. While we have extensive evidence on the prevalence of gender bias in hiring, promotions and wages, we know less about the mechanisms underlying this bias and the extent to which certain personnel policies may mitigate or exacerbate these biases. To test this, I conduct a large scale field experiment with 3,600 employees in 250 schools and randomly vary i). how often managers observe a given employee and ii). whether manager evaluations affect employee's pay or are just used for feedback. First, I find when there are no financial stakes associated with performance evaluations, there is no gender bias. This is true both using data from actual performance evaluations, controlling for the aspects of performance I observe, and for randomized vignettes varying the gender of the teacher. In contrast, when principals' evaluations determine teachers' end of year raise, we see that female teachers receive 10% lower raises, controlling for productivity. However, when principals are randomly assigned to conduct more frequent classroom observations of teachers, this lowers their evaluations of male teachers and results in gender parity in evaluation scores even under financial stakes. Combined this suggests that improving the accuracy of manager information could close the gender gap in performance evaluations, even in high stakes settings.

*I am incredibly thankful for support and advice from Tahir Andrabi, Supreet Kaur, Asim Khwaja, Christopher Walters, and Edward Miguel. David Card, Jishnu Das, Stefano DellaVigna, Frederico Finan, Anne Karing, Samuel Leone, Patrick Kline, Jeremy Magruder, Peter McCrory, Preston Mui, Gautam Rao, Jesse Rothstein, Heather Schofield, and seminar audiences at UC Berkeley provided helpful feedback. I gratefully acknowledge generous funding and support by DFID's RISE Programme, JPAL's Post-Primary Initiative, the Weiss Family Fund, CEGA, the Strandberg Fund, the National Academy of Education/Spencer Dissertation Fellowship, and the Institute for Research on Labor and Employment Fellowship. The 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 my own. I received IRB approval from Pomona College.

†Brown: University of Chicago (christinabrown@uchicago.edu)

1 Introduction

Across the world, women earn half what men earn. This enormous gap in earnings is due to differences in labor force participation, wages and hours, and is significantly worse in many parts of the world. In Pakistan, the setting for this study, men are more than three times as likely to work as women, ranking in the bottom decile for female labor force participation. Even in sectors where women are more prevalent, such as teaching, women earn 70 cents for each dollar men earn. One contributing factor to these gaps is likely the extent of gender bias throughout the personnel process–hiring, promotions, raises, firing, etc, which has been shown to be endemic (Blau and Kahn, 2017; Biasi and Sarsons, 2022).

In addition to concerns about equity, there are also wider productivity consequences of the biased nature of personnel policies. Because many firms and governments worry about biased performance evaluation systems, employees, especially in the public sector, are subject to little to no performance incentives. Take, for example, teachers in the United States. The current system of rigid salary schedules that do not incentivize productivity are a result of lawsuits in the 1930s and 40s demonstrating racial and gender bias in teachers' salaries (Margo, 1990). While we have overwhelming evidence on the prevalence of gender bias throughout the labor market in many settings, we have less understanding about why these disparate outcomes occur and the extent to which other personnel policies amplify or mitigate bias.

In this paper, I use a large-scale experiment with 200 managers and 3,600 teachers across 250 schools in Pakistan to vary features of the personnel system to test the conditions under which gender bias exists and better understand how well-designed HR systems can constrain manager bias. First, I vary whether managers' performance evaluations of employees determine the employee's end of year raise or if they are solely used for feedback. Second, I vary how often managers are required to conduct classroom observations for certain teachers. I then measure teacher productivity in detail, including value-added, daily attendance, time use and measures of pedagogy from classroom observation videos.

First, I show that our sample of managers have relatively minimal overt or taste-based discrimination against female teachers in all three bias measures we conduct. Our managers have significantly more progressive views of women in the workplace than a representative sample of adults in OECD countries. We also do not see any bias when managers are asked to provide sample evaluation scores for example teachers based on a short vignette, where we randomly vary the gender of the teacher in the vignette. Lastly, when we ask teachers about the extent of bias toward or against certain groups, teachers do not believe there is discrimination against or in favor of female teachers.

Given this finding, I construct a conceptual framework in which managers do not exhibit any taste-based discrimination and there are no differences in the mean or variance of productivity by gender. However, there are two key components of the model which result in disparate outcomes by gender: differences in the extent to which employees complain about low evaluation scores (Exley

and Kessler, 2019; Roussille, 2021) and imperfect information about employee quality. These two features result in gender bias in evaluation scores which is amplified when there are larger stakes associated with the evaluation score and when there is less known by managers about employee quality.

Consistent with the framework, I find that when evaluation scores have no financial stakes for employees, male and female teachers receive nearly identical evaluation scores, controlling for teacher productivity. However, when the evaluation score determines the employee's raise, men receive a 10% higher raise than women. This result holds for both male and female managers.

Next, I find that better information helps to close the gender evaluation gap. When managers are required to observe certain teacher's classrooms more frequently, I find this makes them much better able to predict teachers performance on a number of dimensions. Then in turn, I also find that male and female teachers who were observed more frequently receive nearly identical evaluation scores. Whereas, for teachers who were observed less frequently, I find male teachers receive 12% higher raises conditional on the employee's quality. I also find suggestive evidence that the observation treatment is able to mitigate the negative effects of financial stakes, reducing their effects by about two-thirds.

The paper makes two main contributions: first, to our understanding of the mechanisms underlying bias by managers, and second, to demonstrating how certain personnel policies can help mitigate the extent of bias. The paper adds to an extensive list of empirical and theoretical papers demonstrating the extent of gender bias in different aspects of the employment process (Beg et al., 2021; Blau and Kahn, 2017; Grissom and Bartanen, 2022). It extends this literature by showing disparate outcomes by gender are possible even without traditional taste-based or statistical discrimination channels. Instead, I show that even if male and female teachers are equally productive, if their is uncertainty about productivity and male teachers are more likely to advocate or complain for higher scores, then you can end up with biases in performance evaluations.

Second, the paper shows that the specifics of personnel policies can have large effects on the extent of bias. Previous work has demonstrated the existence of bias and has shown that giving managers more discretion in wage-setting results in more gender bias (Biasi and Sarsons, 2022). I build off this work by showing that manager-discretion in and of itself is not necessarily a problem. However, in settings with low information about worker quality, manager discretion increases bias.

2 Conceptual Framework - Gender Bias in Evaluations

In this section, I construct a model to describe wages paid to men versus women in a setting where there are no differences in the mean or variance of employee quality by gender, and there is no taste-based gender discrimination by any managers. The difference in wages by gender and the resulting responsiveness to the HR policy changes we test will arise from two key features of the model: i). managers have imperfect information about worker quality and ii). workers may complain

to their manager about low raises, and, in this model, men are on average more likely to complain, conditional on their productivity. While removing typical statistical discrimination and taste-based discrimination may underestimate the extent of gender discrimination in many labor markets, this model helps demonstrate how differences across groups can arise even in settings where people have "good" intentions.

The performance evaluation system takes place in three stages. First, employees work and produce some output which is observable to the manager. Next, managers provide a performance evaluation score for their employees. Finally, employees may complain to their manager if they are unsatisfied with the score, and higher-level administrators review the scores given out and sanction managers for which there are large discrepancies between worker quality and score. When managers decide on the score to given they take into account these two consequences of their score choice.

Worker Output Employee i produces output y_i, which is the sum of their true ability, θ_i and noise, $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

$$y_i = \theta_i + \epsilon_i \tag{1}$$

Managers observe y_i but not θ_i, and the firm's higher-level administration do not observe y_i.[1]

Manager Chooses Performance Evaluation Score Managers provide a performance evaluation score for employees after observing y_i, and the score is function of both this output and a discretionary component, d_i:

$$s_i = y_i + d_i = \theta + \epsilon_i + d_i \tag{2}$$

Employee and Firm Response to the Score Employees dislike receiving low performance evaluation scores, and will cause disutility to their manager as a result. This could be in the form of complaining to the manager, negotiating for a higher score in the future, lack of cooperation, higher likelihood of turnover, etc. This expected "inconvenience" cost managers face at the time they select the score is:[2]

$$C_i = -c\rho_i s_i \tag{3}$$

where c is the magnitude of the inconvenience cost and ρ_i is the likelihood of complaining for a given employee, conditional on the score received.

[1] We think of y_i as everything that is observable to the manager about the worker's productivity, not just concrete, easily measurable aspects of the production function.

[2] Alternatively, we could think of this as positive utility managers receive from giving better scores to certain individuals. This could be due to favoritism, bias or behaving consistent with social norms.

Finally, the firm would like managers to provide scores which are as close to θ as possible, since $s_i = \theta_i$ is the output-maximizing performance evaluation system. However, they cannot audit every performance evaluation score given, so they audit a subset of scores and punish managers based on the discrepancy between s_i and θ_i.[3] The expectation of the punishment cost at the time of selecting the score is:

$$E[P_i] = p(s_i - \theta_i)^2 = p(\epsilon_i^2 + d_i^2 + 2\epsilon_i d_i) \tag{4}$$

where p is the unit-cost of the punishment. We assume the cost takes this quadratic form in $s_i - \theta_i$ as it is hard to identify small discrepancies in scoring but easier to identify larger ones or blatant manager favoritism. This functional form captures both the increased likelihood of audit and the larger punishment for large discrepancies.

The firm understands that managers do not have perfect information about worker ability at the time of selecting the evaluation score. Therefore, they select the unit-cost of punishment is a function of the noisiness of the production function, such that $p = \frac{1}{\sigma_\epsilon^2}$. This holds costs of the punishment strategy fixed as σ_ϵ^2 changes.[4]

Manager's Choice of Evaluation Score Managers select the discretionary component of the salary to minimize the inconvenience and punishment costs they expect to face in the next period[5]:

$$\begin{aligned} u(d_i) &= \min_{d_i} E[-c\rho_i s_i + p(s_i - \theta_i)^2] \qquad (5) \\ &= \min_{d_i} E[-c\rho_i(\theta_i + \epsilon_i + d_i) + \frac{1}{\sigma_\epsilon^2}(\epsilon_i + d_i)^2] \\ \frac{\partial u_i}{\partial d_i} &= E[-c\rho_i + 2\frac{1}{\sigma_\epsilon^2}(\epsilon_i + d_i)] = 0 \\ d_i^* &= \frac{c\rho_i \sigma_\epsilon^2}{2} \qquad (6) \end{aligned}$$

Therefore an individual's evaluation score will be $s_i^* = y_i + \frac{c\rho_i\sigma_\epsilon^2}{2}$.

Gender Gap in Evaluation Score Men and women's ability is draw from the same distribution $\Theta \sim \mathcal{N}(\mu, \sigma_\theta^2)$, so both the mean and variance in ability are the same between men and women. The noisiness in the observation of output, ϵ is also drawn from the same distribution. Consistent with the literature on bargaining and negotiation, we assume that men are more likely to contest a low evaluation score, so $\rho_m > \rho_f$. The difference in expected scores, conditional on ability, by gender

[3] Firms are able to unearth θ through this costly audit process but do not know exactly what information managers knew at the time of scoring, y.

[4] The firm sees the benefits of the punishment system as $E[pd_i^2]$ (the extent to which is catches large discretion by the manager) and the costs as $E[p(\epsilon_i^2 + 2\epsilon_i d_i)]$ (the manager being unnecessarily punished for the noisiness of the production function).

[5] Managers have correct expectations about ρ_i, p and c.

then is:

$$\frac{\partial s_i^*}{\partial female}|_{\theta_i} = y_f^* - y_m^* + \frac{c\sigma_\epsilon^2}{2}(\rho_f - \rho_m) \tag{7}$$
$$= \frac{c\sigma_\epsilon^2}{2}(\rho_f - \rho_m) < 0$$

Effect of Noise and Financial Stakes on Gender Gap The key comparative statics we are interested in are how scores (and the gender gap in scores) respond to changes in the accuracy of information managers have (σ_ϵ^2) and the magnitude of the inconvenience cost (c):

$$\frac{\partial s_i^*}{\partial c} = \frac{\rho_i \sigma_\epsilon^2}{2} > 0 \qquad \frac{\partial s_i^*}{\partial \sigma_\epsilon^2} = \frac{c\rho_i}{2} > 0 \tag{8}$$

As we would expect, the score is increasing in the inconvenience cost and noisiness of the production function. These parameters also affect the size of the gender gap:

$$\textit{Prediction 1:} \frac{\partial^2 s_i^*}{\partial c \partial female}|_{\theta_i} = \frac{\sigma_\epsilon^2}{2}(\rho_f - \rho_m) < 0 \tag{9}$$

$$\textit{Prediction 2:} \frac{\partial^2 s_i^*}{\partial \sigma_\epsilon^2 \partial female}|_{\theta_i} = \frac{c}{2}(\rho_f - \rho_m) < 0 \tag{10}$$

$$\textit{Prediction 3:} \frac{\partial^3 s_i^*}{\partial c \partial \sigma_\epsilon^2 \partial female}|_{\theta_i} = \frac{1}{2}(\rho_f - \rho_m) < 0 \tag{11}$$

Increases in the noisiness of the production function, the costliness of employee complaints and their interaction increase the size of the gender gap.

3 Experimental Design

3.1 Mapping the Model to an Experiment

To test predictions 1-3, we need variation in the magnitude of the inconvenience cost (c) and the noisiness between what managers observe and teacher's actual ability (σ_ϵ^2). In order to vary the former, we will introduce financial stakes for the employee associated with the manager evaluation score, tying the score to the employee's raise. From conversations with managers, this makes the evaluation score significantly more consequential for employees and results in increased push back against managers for poor scores. To vary noise, we will introduce variation in how frequently managers conduct in person observations of employee effort. These two treatments are discussed in detail in section 3.2.

Our predictions also rely on being able to measure employees true ability (θ) or the information managers have about employee's true ability (y). To capture this we measure a wide variety of teacher and student outcomes: value-added on test scores, classroom observation pedagogy scores,

daily attendance, time use outside of teaching, and intrinsic motivation. In section 3.4, we discuss each measure.

3.2 HR Policy Interventions

In order to test our hypotheses, we introduce two variations in how managers within the school system evaluate their employees. The study was conducted from October 2017 to June 2019 with a private school chain that operates nearly 300 schools located across Pakistan. Figure 1 presents the timeline of interventions and data collection activities.

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

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

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

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

Once scores are finalized, managers sit down with the employee to discuss their score and provide feedback on the performance in the previous year. They also generally discuss areas to work on

improvement in the future. Most employees find the performance evaluation process helpful and constructive.

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

- **Control:** In control schools, managers complete the performance evaluation cycle as described above. Employee's end of year raise is then determined one of two ways:
  - *Flat Raise:* Employees receive a raise of 5% of their base salary
  - *Objective Raise:* Teachers receive a raise from 0-10% based on their within-school percentile value-added (Barlevy and Neal, 2012) averaged across all students they taught during the spring and fall term exams.[6]
- **Treatment: Subjective Raise:** Teachers receive a raise from 0-10% based on their performance evaluation score.

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

Randomization was conducted at the school level, so all teachers at the school were under the same type of raise system.[7,8] The contract applied to all core teachers (those teaching Math, Science, English, Urdu, and Social Studies) in grades 4-13. Elective teachers and those teaching younger grades received the status quo contract. All three contracts have equivalent budgetary implications for the school. I over-sampled the number of subjective treatment arm schools due to partner requests, so the ratio of schools is 4:1:1 for subjective treatment, objective treatment, and control, respectively.

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

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

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

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

3.3 Data

I draw on four sources of data: i). school system administrative data, ii). student test scores, iii). manager and employee surveys, and iv). classroom observations in order to measure the following categories of outcomes:

Type	N	Source	Outcomes
Teachers			
"Ground truth"	1,500	Class video	Rubric covering 20 aspects of pedagogy (Araujo et al, 2016)
	3,600	Admin data	Value-added (From 5 years of student test scores)
	9,100	Admin data	Daily clock in and out time
Managers			
Beliefs	189	Survey	Rate teachers on several criteria
	189	Admin data	Rate teachers on several criteria (after observation)
Preferences	189	Survey	Vignettes (rating hypothetical teachers)
	189	Survey	Rank importance of teacher behaviors
	189	Evaluation	Points allocated to criteria
Evaluation	189	Evaluation	Total score and criteria-level score
Bias	189	Survey	World Values Survey questions
	189	Survey	Teacher's rating of manager's bias
	189	Survey	"Audit": Varying gender of name in vignette

Administrative data The administrative data details position, salary, performance review score, attendance, and demographics for all employees. We also have biometric clock in/out data for all

schools. The data was provided by the school system for the period of July 2016 to June 2019. It includes classes and subjects taught for all teachers, and end of term standardized exam scores for all students (linked to teachers). From September through December 2018, we also have data on classroom observations conducted by managers. Managers use a similar rubric to the one used by the research team to conduct classroom observations (detailed below).

Student test scores 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]

Manager and employee surveys At baseline and endline, we measure teacher's contract preferences, beliefs about their value-added, and risk preferences. We also conduct a time use survey to understand how much time teachers spend on lesson planning, helping with administrative tasks. The survey also included measures of intrinsic motivation (Ashraf et al., 2020), efficacy (Burrell, 1994), and checks on what teachers understood about their assigned contract. The endline survey was conducted online with teachers and managers in spring and summer 2019. Appendix table B3 lists the survey items used for each area along with their source. The manager endline survey measured managers' beliefs about teacher quality and measured management quality using the World Management Survey school questionnaire.[10] The endline survey was conducted online with teachers and managers in spring and summer 2019. 6,080 teachers and 189 managers were surveyed.

Classroom observations 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.[11,12] We also

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

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

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

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

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.

3.4 Measuring Employee Quality

Because most of our analysis will look at whether managers are rating female and male employees differentially, I control for teacher quality and effort using comprehensive data on numerous outcomes. This helps to assuage concerns that evaluations are different due to differences in female versus male teacher performance.

Value-added To measure teacher's "ability", θ, we calculate teacher value-added (VA) using student test scores from June 2016 and 2017, the two years prior to the randomized controlled trial. This allows us to measure teacher effectiveness in the absence of the treatments. We follow Kane and Staiger (2008) in constructing empirical Bayes estimates of teacher value-added. Appendix Appendix A.A provides details on the calculation used.

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

to achieve an IRR of 0.7 with the practice data. 10% of videos were also double reviewed to ensure a high level of IRR throughout the review process. We have a high degree of confidence in the internal reliability of the classroom observation data, but because this was not conducted by the Teachstone staff, we caution against comparing these CLASS scores to CLASS data from other studies.

In-class effort I measure teacher's in-class effort using the classroom observation data. All results control for the teacher's score along the 12 dimensions of the CLASS rubric: positive climate, teacher sensitivity, regard for student perspectives, behavioral management, productivity, negative climate, instructional learning formats, content understanding, analysis and inquiry, quality of feedback, instructional dialogue and student engagement. I also control for four additional dimensions which some managers have used in their evaluations: use of English (except in Urdu classes), use of technology, displays on the classroom walls and time spent on test preparation.

Out of class effort To measure teacher's effort outside the classroom, I rely on administrative data which tracks teachers' attendance and teacher's self report from the endline survey about time spent on: teaching, lesson planning, grading, administrative tasks, afterschool tutoring, interacting with parents, interacting with fellow teachers and interacting with their manager.

Other teacher qualities Finally, I control for several aspects of the teacher's personality which may affect the way they contribute to the school beyond the effect on their own students. These measures come from self-report during the baseline survey and measure teachers intrinsic motivation, efficacy, and long-term career plans. Lastly, I control for teacher experience.

3.5 Sample and Intervention Fidelity

Employees The study was conducted with a large, high fee private school system in Pakistan. The student body is from an upper middle-class and upper-class background. School fees are $900 USD. Table 2, panel A, presents summary statistics for our sample teachers compared to a representative sample of teachers in Punjab, Pakistan (Bau and Das, 2020). Our sample is mostly female (81%), young (35 years on average), and the median experience level is 10 years, but a quarter of teachers are in their first year teaching. Nearly all teachers have a BA, and 68% have some post-BA credential or degree. Teachers are generally younger and less experienced than their counterparts in public schools, though they have more education.

Managers Managers here are either a principal in small schools or a vice principal in larger schools. They are tasked with overseeing the overall operations of the school and managing employees, including teachers and other support staff. Table 3 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, managers in our sample spend much more time directly observing teachers. They do about twice the number of classroom observations each year (4.7 versus 2.5 in the US). They also rate themselves higher in most areas of the management survey questions (4.3 versus 2.8 out of 5), including formal evaluation, monitoring, and feedback systems for teachers. This is an important difference as these management practices could positively effect the success of the subjective treatment arm, and may help us understand the extent of external validity of these results.

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

Schools under the various performance evaluation treatments appear to be balanced along baseline covariates. Appendix table B1 compares schools along numerous student and teacher baseline characteristics. Of 27 tests, one is statistically significant at the 10% level, and one is statistically significant at the 5% level, no more than we would expect by random chance. Results control for these few unbalanced variables.

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

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

4 Results

In the following section, we will look at the effect of two changes to the performance evaluation process: delinking performance evaluations from financial compensation for employees and increasing the time managers spend observing workers. For each, we will show the effects of the intervention on the extent of gender bias in evaluation scores. Finally, we will see if these effects vary by manager characteristic, such as gender, age and extent of bias (as measured by survey questions).

4.1 Measured Gender Bias by Managers

We measure gender attitudes and more overt forms of gender bias using three sources of data: i). managers' responses to survey questions designed to assess gender attitudes, ii). managers' hypothetical evaluation scores for a series of vignettes, which vary the gender of the teacher described in the vignette, and iii). teachers' beliefs about the extent of gender bias by their manager.

Manager Survey Responses First, during the endline survey, we ask managers to agree or disagree with several statements used in the World Values Survey to gauge beliefs around women in the workplace. Figure 3 shows the distribution of responses for our managers relative to the average respondents in OECD countries and South Asian countries which participated in the 7th wave of the World Values Survey. On all three statements, our sample rates as more gender progressive than the average respondent in OECD countries. These patterns hold for both the male and female managers in our sample. South Asia, in general, though, rates significantly more conservative and is the most gender conservative region based on the World Values Survey responses.

Manager Ratings of Vignettes However, we may be concerned that individuals do not feel comfortable sharing their views truthfully in the survey as they know it is conducted by a foreign research team. In order to perform a more naturalistic task that is less blatantly about gender, we provide a series of vignettes to managers about hypothetical employees and ask them to give a performance evaluation score of that employee. In the vignette, we randomly vary the teacher's name (traditionally male versus traditionally female) and the description the teacher's percentile rank in terms of value-added, behavioral management and attendance. The task is framed to the managers as trying to gauge what aspects of teaching managers value, with nothing about gender mentioned during the task. Managers repeat the task for several vignettes with different characteristics.

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

We may be concerned that this approach using vignettes simply does not have any predictive power. This could be the case if, for example, teachers are reading the survey quickly and do not even notice the name listed. To test this, we can see how female vignettes are rated for managers who gave more regressive answers to the World Values Survey. Table 6 shows the same specification as in table 5, column (2) but includes an interaction with various manager characteristic's such as the

manager's gender, age, extent of bias on the World Values Survey questions. We find that managers who have higher levels of gender bias on these questions are more likely to rate vignettes with female names lower. In particular, responding that you agree or strongly agree that "In general, it is better for a family if a woman has the main responsibility for taking care of the home and children rather than a man." is associated with lower scores for female-named teachers. This helps assuage concern that our measure was not sensitive enough to pick up gender bias.

Teacher's Beliefs about Manager Bias Finally, what may be most pertinent is the extent to which teachers believe their manager is biased toward certain groups. To test this we ask teachers the following question: "It can be easy for our own biases and subjectivity to get in the way of making objective appraisals. How much bias do you think your line manager has when they conduct appraisals of... [new/female/older] teachers", with choices from "lots of bias against" through "lots of bias towards". Figure 5 shows teachers average responses for schools who were part of the financial stakes treatment versus those that were not. We find that on average teachers do not believe there is any bias against or in favor of female teachers, and this is true irrespective of treatment status. Though teachers do believe there is slight bias in favor of older teachers.

Combined, these three pieces of evidence suggest, on the face of it, our sample of managers does not hold discriminatory views of women in the workplace, except for a small fraction of managers. This finding informs why the conceptual framework does not include taste-based gender discrimination. However, as we will see in the next section, consistent with the framework, even without more overt kinds of discrimination by managers, we can still end up with disparate employment outcomes by gender.

4.2 Effect of Financial Stakes on Bias

Under the financial stakes treatment, teachers end of year raise is determined by their manager ("subjective"), as compared to being determined based on their students' test scores ("objective") or everyone receiving the same raise irrespecitve of performance ("flat"). To first show that teachers understood and took seriously this policy change, figure 6 shows the effect of the subjective raise treatment relative to the flat raise on student test scores and teacher behavior. We find that in response to the treatment students have higher test scores, teachers are more likely to tailor their lesson to address different students' needs and teachers are more likely to show up for work.

To test prediction 1, whether increasing the costs associated with a low evaluation score affects the extent of gender bias, we compare teachers' raises in the subjective treatment versus those in

schools without a financial stake (objective and flat pooled). Our main specification is:

$$PredictedRaise_{is} = \beta_0 + \beta_1 FinancialTreatment_s + \beta_2 Female_i + \beta_3 FinancialTreatment_s * Female_i + \chi_i + \epsilon_{is} \quad (12)$$

where the dependent variable is the teachers' predicted end of year raise using their manager-assigned performance evaluation score. Note, that in the financial stakes this is the actual raise paid out to teachers and in the control, their actual raise was either a flat 5% of their current salary or was based on their students' test scores. $FinancialTreatment_s$ is a dummy for whether the employee's school, s, had financial stakes tied to the raise, and $Female_i$ is a dummy for whether the employee is female. χ_i are controls for teacher's quality: value-added, 16 dimensions of classroom quality, attendance, time use outside the classroom and intrinsic motivation. β_2 tells us if female teachers receive a different evaluation than male teachers, with similar performance, under no financial stakes. The main coefficient of interest is β_3 which tests whether men and women's scores are differentially affected by tying them to financial stakes. Standard errors are clustered at the school level (the unit of randomization).

I find when their are no financial stakes of the performance evaluation male and female teachers receive nearly identical evaluations, controlling for everything we can measure about teacher productivity. However, when there is a financial stake, women receive significantly lower evaluation scores, equivalent to a 10% lower raise. Figure 7 and table 7 present the results of eq. 12. In the figure, each bar plots the average effective raise based on evaluation score, by treatment status and gender, controlling for teacher productivity. The first two bars show men and women receive similar scores in control schools, with an average difference of \$2, off a mean raise of \$364. The next two bars show there is a significant difference in raise received by female teachers in the financial stakes treatment schools. Men receive an additional \$37 ($p < 0.00$) compared to female teachers with similar measured productivity. In addition, the interaction between the financial treatment and gender (β_3) is \$34 ($p = 0.04$). The results are also similar, if we do not include any controls for teacher productivity, β_3= \$30 ($p = 0.08$), and are shown in table B2.

4.3 Effect of Information on Bias

For the second treatment arm, we will first show that having managers do more classroom observations does actually improve the information they have about teacher quality. Table 8 and figure 8 shows the relationship between managers' beliefs about different aspects of teacher performance and their actual performance. Column (1) pools across all four aspects of teacher quality (attendance, disciplinary management of students, focus on analysis/inquiry skills and value-added) and columns (2) -(5) presents each of these components separately. We can see that, on average, managers seem to have fairly accurate information about teacher attendance and disciplinary management but are less accurate about the other aspects of teacher performance.

Finally, column (6) shows the interaction between whether the teacher was assigned to be observed more frequently. We find that managers are about twice as accurate in evaluating teacher performance when they were required to observe them more frequently. This suggests that the treatment worked in improving the accuracy of information managers have about their employees.

To test prediction 2, whether improving manager information about worker quality decreases bias, we compare teachers under teh classroom observation treatment versus those in the stas quo. Our main specification is:

$$\begin{aligned} PredictedRaise_{is} = \beta_0 + \beta_1 ObservationTreatment_i + \beta_2 Female_i \\ + \beta_3 ObservationTreatment_i * Female_i + \chi_i + \epsilon_{is} \end{aligned} \tag{13}$$

where the dependent variable is the teachers' predicted end of year raise using their manager-assigned performance evaluation score. $ObservationTreatment_s$ is a dummy for whether the employee was assigned to be observed more frequently by their manager, and $Female_i$ is a dummy for whether the employee is female. χ_i are controls for teacher's quality: value-added, 16 dimensions of classroom quality, attendance, time use outside the classroom and intrinsic motivation. β_2 tells us if female teachers receive a different evaluation than male teachers, with similar performance, under the status quo level of observations. The main coefficient of interest is β_3 which tests whether men and women's scores are differentially affected by increased classroom observations. Standard errors are clustered at the teacher level (the unit of randomization).

I find that under the status quo level of manager observation, female teachers receive significantly lower raises, controlling for teacher productivity. However, when managers are required to observe teachers more frequently, the gender gap is completely closed. Figure **??** and table 7 present the results of eq. 13. In the figure, each bar plots the average effective raise based on evaluation score, by treatment status and gender, controlling for teacher productivity. The first two bars show men receive significantly higher scores under the status quo level of observations, with an average difference of \$42 (12% of their raise, $p < 0.00$) . The next two bars show that under the additional monitoring by managers, male and female teachers now receive nearly identical raises, with an average difference of \$0.40. In addition, the interaction between the observation treatment and gender (β_3) is \$42 ($p = 0.01$). The results are also similar, if we do not include any controls for teacher productivity, β_3= \$31 ($p = 0.09$), and are shown in table B2.

Finally, to test prediction 3, I test for an interaction between the two treatments to see the extent to which information may mitigate some of the detrimental effects of financial stakes. The

specification I test is:

$$\begin{aligned} PredictedRaise_{is} = \beta_0 &+ \beta_1 ObservationTreatment_i + \beta_2 Female_i + \beta_3 ObservationTreatment_i * Female_i \\ &+ \beta_4 FinancialTreatment_s + \beta_5 FinancialTreatment_s * Female_i \\ &+ \beta_6 FinancialTreatment_s * Observation\ \ Treatment_i \\ &+ \beta_7 FinancialTreatment_s * Observation\ \ Treatment_i * Female_i \\ &+ \chi_i + \epsilon_{is} \end{aligned} \tag{14}$$

where the dependent variable, treatments and controls are the same as in eq. 12 and 13. β_3 tells us the effect of the observation treatment on gender bias in the finance control schools. β_5 tells us the effect of the finance treatment on gender bias for the observation status quo teachers. β_7 tells us whether there is a compound effect of the two treatments.

I find some slight evidence that the financial treatment is partly mitigated under the observation treatment. Table 7, col. 4, and figure 10 show the interaction of the two treatments. Under the observation status quo, the financial treatment causes female teachers to receive a \$51 lower raise (14% of their total raise, $p < 0.05$). However, under the observation treatment the effect of the financial treatment is only \$13, though the interaction terms β_7 is not statistically significant.

4.4 Heterogeneity by Managers

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

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

Overall, we do not find a dramatic difference in the treatment effect by manager characteristic, though our standard errors are large, so we cannot reject relatively large effects in either direction. Table 9 and table 10 present results from eq. 12 and eq. 13 adding in an interaction with the respective manager covariate. One suggestive pattern we see is that the financial stakes actually have less of a negative effect on female ratings for managers who are more "biased" as measured

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

5 Conclusion

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

6 References

References

Ashraf, Nava, Oriana Bandiera, Edward Davenport, and Scott S. Lee, "Losing Prosociality in the Quest for Talent? Sorting, Selection, and Productivity in the Delivery of Public Services," *American Economic Review*, May 2020, *110* (5), 1355–1394.

Azam, Mehtabul and Geeta Kingdon, "Assessing Teacher Quality in India," *Working Paper*, October 2014, p. 31.

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

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

Beg, Sabrin, Anne Fitzpatrick, and Adrienne M. Lucas, "Gender Bias in Assessments of Teacher Performance," *AEA Papers and Proceedings*, May 2021, *111*, 190–195.

Biasi, Barbara and Heather Sarsons, "Flexible Wages, Bargaining, and the Gender Gap," *Quarterly Journal of Economics*, January 2022, *137* (1), 215–266.

Blau, Francine D. and Lawrence M. Kahn, "The Gender Wage Gap: Extent, Trends, and Explanations," *Journal of Economic Literature*, September 2017, *55* (3), 789–865.

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

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

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

Exley, Christine and Judd Kessler, "The Gender Gap in Self-Promotion," Technical Report w26345, National Bureau of Economic Research, Cambridge, MA October 2019.

Grissom, Jason A. and Brendan Bartanen, "Potential Race and Gender Biases in High-Stakes Teacher Observations," *Journal of Policy Analysis and Management*, December 2022, *41* (1), 131–161.

Haerpfer, Christian, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen, "World Values Survey Wave 7 (2017-2020) Cross-National Data-Set," 2021. Type: dataset.

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

Margo, Robert A., *Race and schooling in the South, 1880-1950: an economic history* Long-term factors in economic development, Chicago: University of Chicago Press, 1990.

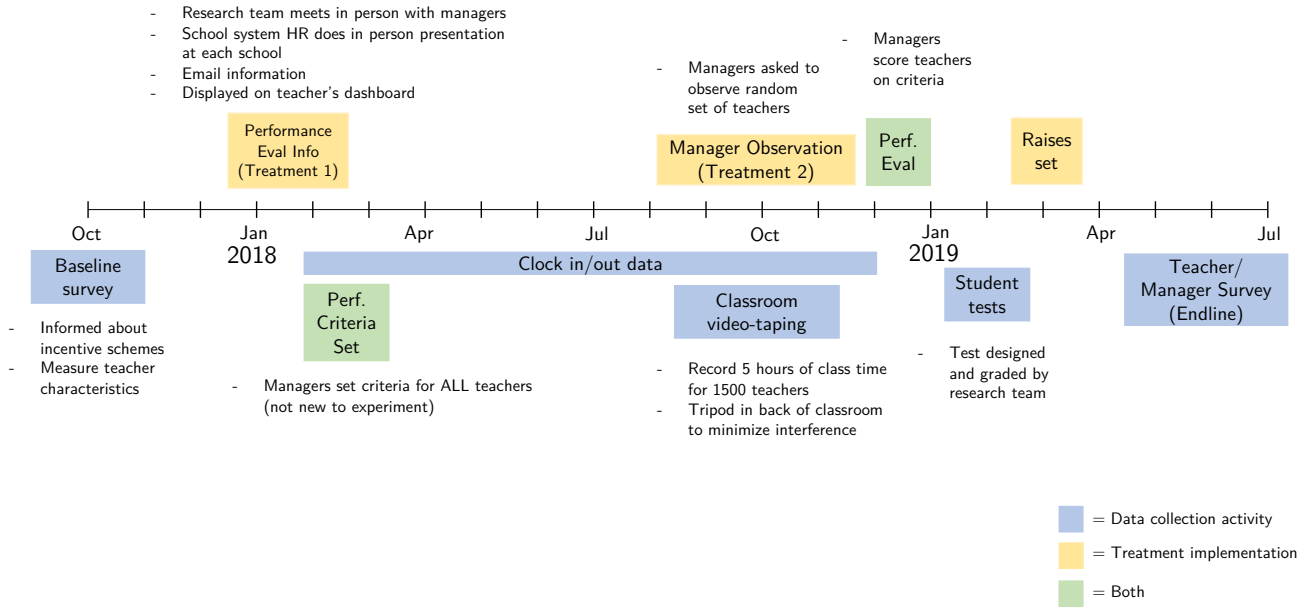
National Center for Education Statistics, *Schools and Staffing Survey, 2010-2011: [United States]*, U.S. Dept. of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 2011.

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

Roussille, Nina, "THE CENTRAL ROLE OF THE ASK GAP IN GENDER PAY INEQUALITY," 2021, p. 99.

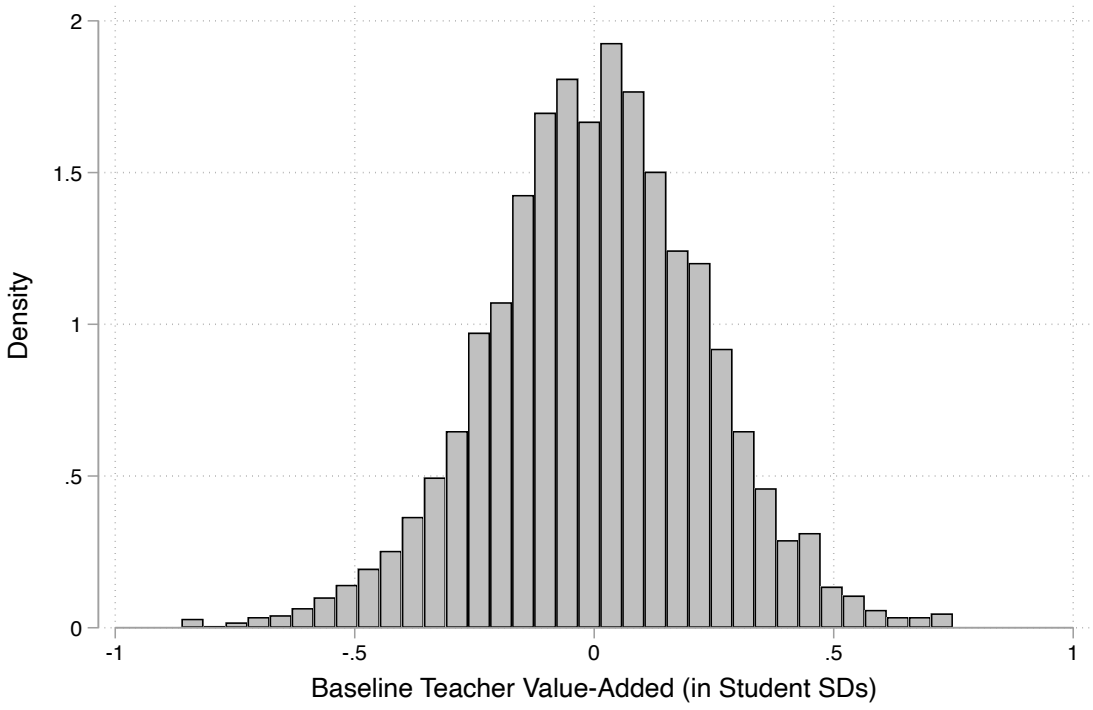
7 Figures

Figure 1: Experimental Timeline



Notes: This figure shows the timeline of the experiment from October 2017 through July 2019 and includes treatment implementation activities and data collection activities/periods.

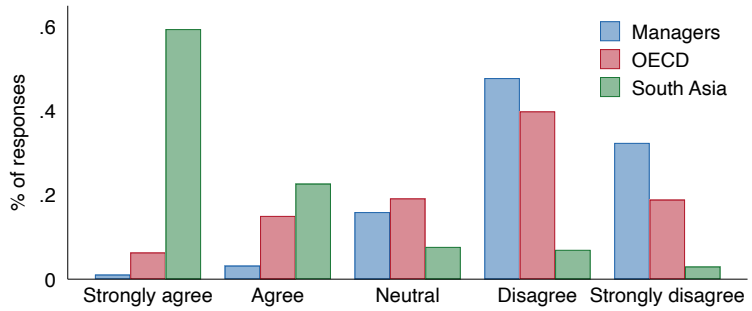
Figure 2: Distribution of Teacher Value-Added at Baseline



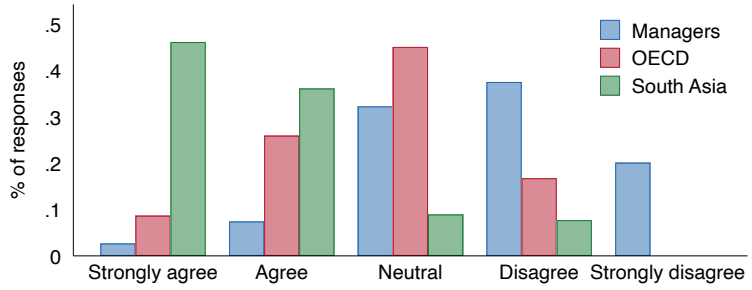
Notes: This figure presents the distribution of teacher value-added for 3,687 teachers in the school system at baseline. Teacher value-added is calculated using administrative test score data from June 2016 and June 2017 (the two years prior to the intervention). Estimates are calculated following Kane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

Figure 3: Managers' Stated Beliefs about Women and Employment

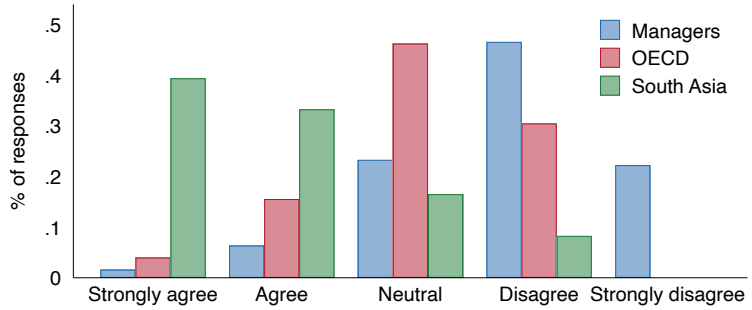
Panel A: When jobs are scare, men should have more rights to a job than women



*Panel B: It is better for a family if a women has the main responsibility for taking care of a home**



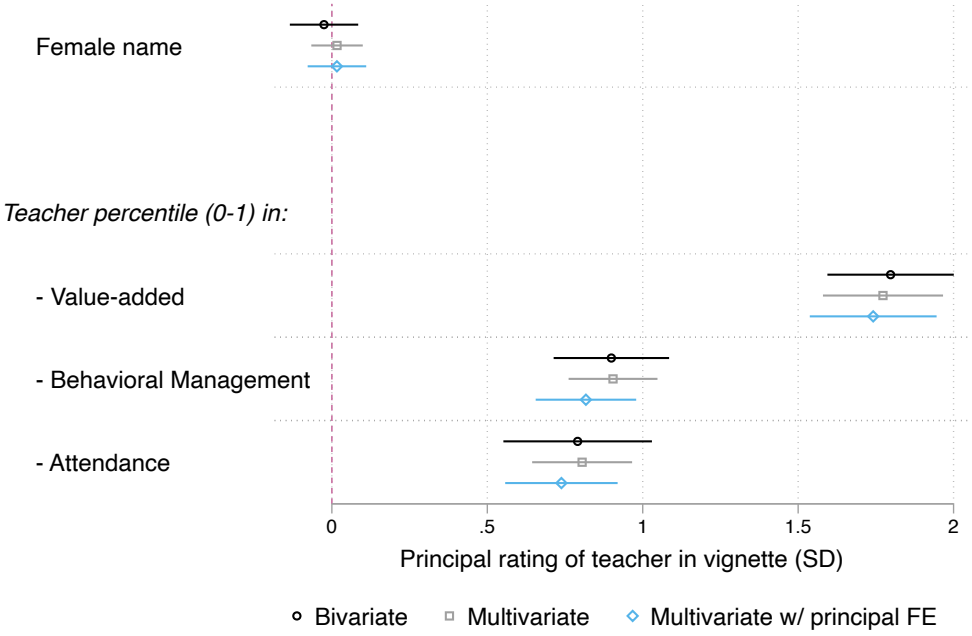
*Panel C: Men are better suited than women to each math and science***



Notes: This figure shows the distribution of responses to survey questions about the role of women in the workforce for our sample of managers and for a representative sample of individuals in several OECD countries and several South Asian countries.

- The y axis shows the percent of respondents who selected the given choice in response to the statement.
- The data for *Managers* comes from the endline survey conducted in spring 2019, conducted with 189 principals and vice principals from our school study sample.
- Data for the *OECD* and *South Asia* sample come from the World Values Survey Wave 7 (2017-2021) (Haerpfer et al., 2021). The World Values Survey is an in-person survey conducted with a representative sample of the adult population from 50 countries. This OECD countries in this wave include Australia, Chile, Colombia, Germany, Greece, Japan, Mexico, New Zealand, Turkey and the United States, with a total sample of 15,598. The South Asian countries in this wave are Bangladesh and Pakistan, with a total sample of 3,111.
- To make the question more relevant to the study sample, we made some changes to the statements in Panel B and C. The statements listed above each figure is the one used in the study survey.
  *In Panel B, the corresponding WVS question is: "When a mother works for pay, the children suffer".
  **In Panel C, the corresponding WVS question is: "On the whole, men make better business executives than women do"

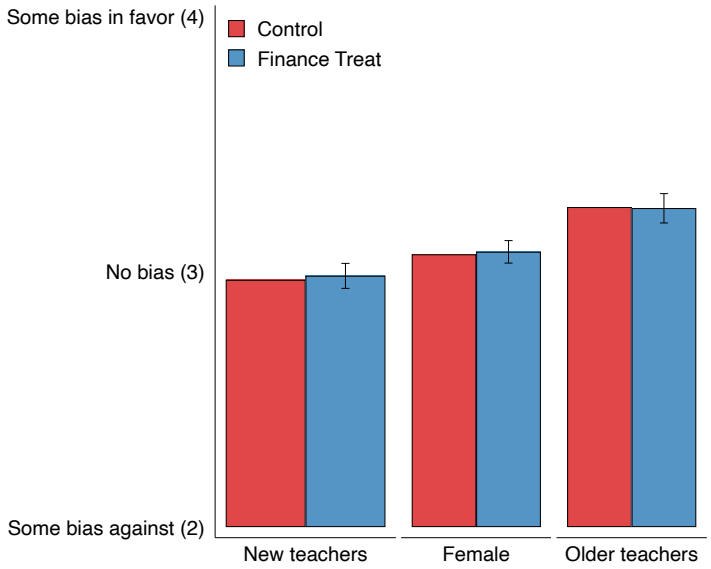
Figure 4: Manager Evaluation Score of Example Vignettes



Notes: This figure shows the effect of different vignette characteristics on the hypothetical evaluation score provided by managers.

- The x-axis shows the magnitude of the coefficients in standard deviations of evaluation score and the y-axis shows the regressors.
- Data is from the endline survey conducted with 189 managers.

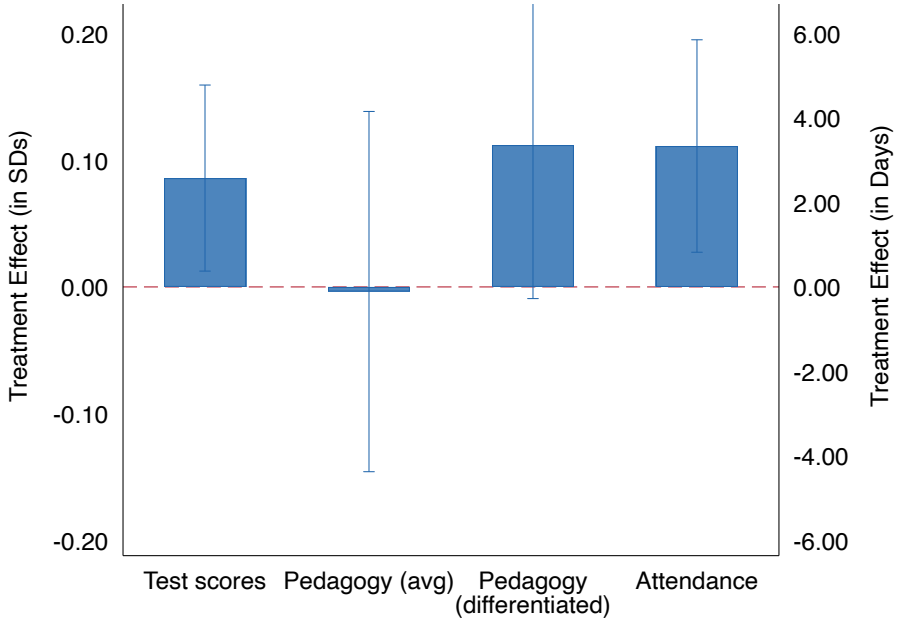
Figure 5: Teachers' Perceptions of Bias in Evaluation Scores by Treatment



Notes: This figure shows teachers' beliefs about the extent of bias toward certain groups by the schools treatment status

- Data comes from the endline employee survey with 5,248 respondents.
- The survey question stated: "It can be easy for our own biases and subjectivity to get in the way of making objective appraisals. How much bias do you think your line manager has when they conduct appraisals of...[new/female/older] teachers". The choice options are: Lots of bias against (1), Some bias against (2), No bias (3), Some bias in favor (4), Lots of bias in favor (5).
- The y axis is the average teacher response to the question. The red bars are for teachers for whom the evaluation score did not have a financial stake, and the blue bars are for teachers in which the evaluation score determined their raise.

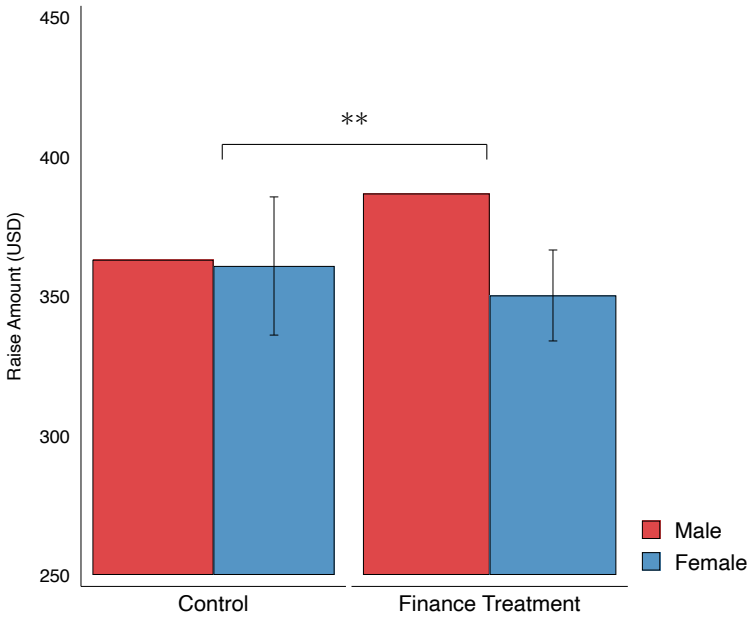
Figure 6: Effect of Financial Stakes on Student Outcomes and Teacher Behavior



Notes: This figure shows the effect of the financial stakes treatment relative to the flat raise condition on teacher behavior and student outcomes.

- The y axis plots the treatment effect coefficient of the subjective performance raise treatment (financial stakes) relative to the flat raise treatment (no financial stakes) with 95% confidence intervals. The left axis shows treatment effects in standard deviations and applies to the first three outcomes: *test scores, pedagogy (avg)* and *pedagogy (differentiated)*. The right axis shows treatment effects in days and applies to the right-most outcome: *attendance*.
- The outcome *test scores* comes from endline student tests conducted in January 2019 with 40,500 students. *Pedagogy (avg)* and *pedagogy (differentiated)* are from the classroom observations conducted in fall 2018 with 1,750 teachers. The first is the average score across all 12 dimensions of the CLASS rubric (Pianta et al., 2012), and the latter restricts to the dimensions of the rubric which measure the extent to which the teacher tailors the lesson to different student's needs. *Attendance* comes from administrative biometric clock in data for 6,390 teachers. Observations are at the student-test level, classroom observation level, and teacher-day level, respectively.
- Standard errors are clustered at the school level, the unit of randomization. All regressions control for randomization strata, the test score results control for baseline test performance, and the classroom observation results control for observer fixed effects.

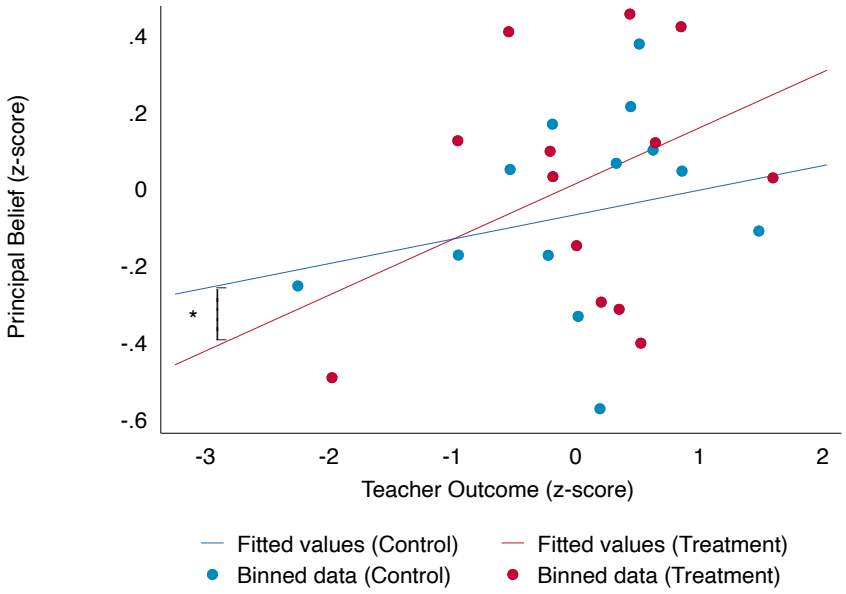
Figure 7: Effect of Financial Stakes Treatment on Gender Bias



Notes: This figure shows the average evaluation score (in terms of the effective raise amount the score would correspond to) for teachers by gender and treatment status.

- The y axis is the average evaluation score in terms of the effective raise that score would correspond to in USD. The first two bars show the average raise by gender for teachers in the control schools (where teachers raise was not based on manager evaluation). The second two bars show the average raise by gender for teachers in the financial stakes schools. The bracket above the bars shows the statistical significance of the interaction between treatment and gender.
- Data is from evaluation scores from December 2018.
- Standard errors are clustered at the school level (the unit of randomization). All regressions control for teacher value-added, classroom observation score in each of the 12 dimensions of the CLASS rubric (Pianta et al., 2012), attendance, self-reported time use, intrinsic motivation and locus of control. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

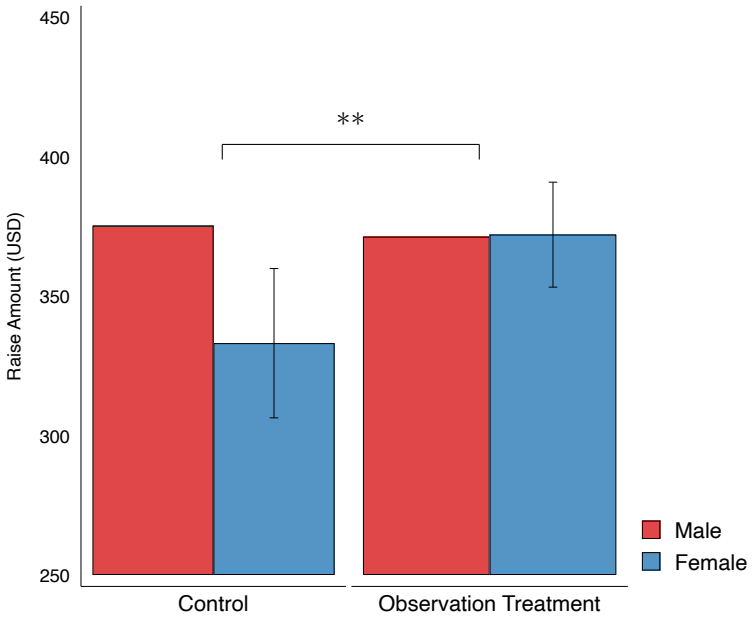
Figure 8: Effect of Observation Treatment on Principal Beliefs



Notes: This figure shows the relationship between teacher quality and manager prediction of quality by observation treatment status.

- The y axis is manager's rating of teacher quality, and the x axis is the teacher's measured productivity. The blue values are for the status quo level of observation and the red values are for teachers who were under the frequent observation treatment.
- Observations are at teacher-outcome level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

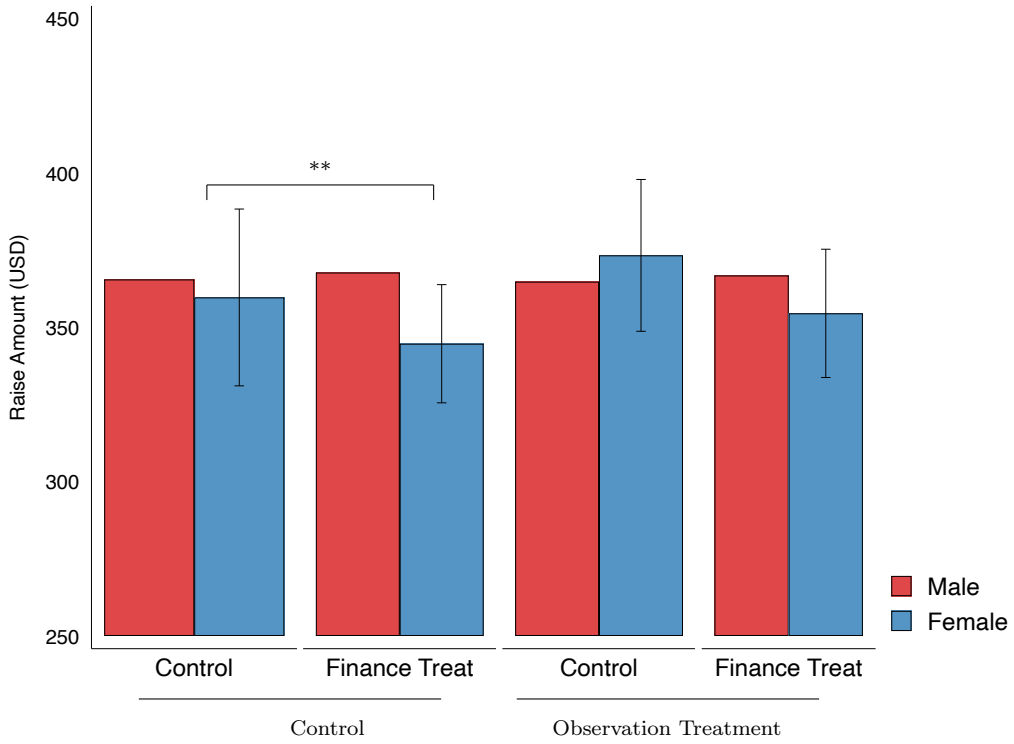
Figure 9: Effect of Employee Observations on Gender Bias



Notes: This figure shows the average evaluation score (in terms of the effective raise amount the score would correspond to) for teachers by gender and treatment status.

- The y axis is the average evaluation score in terms of the effective raise that score would correspond to in USD. The first two bars show the average raise by gender for teachers under the status quo level of classroom observations. The second two bars show the average raise by gender for teachers in frequent observations treatment. The bracket above the bars shows the statistical significance of the interaction between treatment and gender.
- Data is from evaluation scores from December 2018.
- Standard errors are clustered at the teacher level (the unit of randomization). All regressions control for teacher value-added, classroom observation score in each of the 12 dimensions of the CLASS rubric (Pianta et al., 2012), attendance, self-reported time use, intrinsic motivation and locus of control. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Figure 10: Effect of Financial Stakes and Employee Observations on Gender Bias



Notes: This figure shows the average evaluation score (in terms of the effective raise amount the score would correspond to) for teachers by gender and treatment status.

- The y axis is the average evaluation score in terms of the effective raise that score would correspond to in USD. The first, second, fifth and sixth bars show the average raise by gender for teachers in the control schools (where teachers raise was not based on manager evaluation). The second third fourth, seventh and eighth bars show the average raise by gender for teachers in the financial stakes schools.
- The first four bars show the average raise by gender for teachers under the status quo level of classroom observations. The latter four bars show the average raise by gender for teachers in frequent observations treatment. The bracket above the bars shows the statistical significance of the interaction between treatment and gender.
- Data is from evaluation scores from December 2018.
- Standard errors are clustered at the school level (the unit of randomization). All regressions control for teacher value-added, classroom observation score in each of the 12 dimensions of the CLASS rubric (Pianta et al., 2012), attendance, self-reported time use, intrinsic motivation and locus of control. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

8 Tables

Table 1: World Values Survey Summary Statistics

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

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

Table 2: Descriptive Statistics about Study Sample and Comparison Sample

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

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

Table 3: Descriptive Statistics about Mangers in Study and Comparison Sample

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

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

Table 4: Principal Beliefs about Teacher Quality

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

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

Table 5: Manager Rating by Vignette Characteristic

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

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

Table 6: Manager Rating by Vignette and Manager Characteristic

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

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

Table 7: Raise Amount by Treatment and Gender

	Predicted Raise Amount (USD)			
	(1)	(2)	(3)	(4)
Female	-28.21***	-2.228	-41.90***	-18.43
	(6.600)	(12.59)	(13.15)	(18.71)
Financial Treatment		23.36		36.88
		(20.81)		(28.91)
Financial Treatment*Female		-34.33**		-51.01**
		(15.32)		(25.65)
Observation Treatment			-36.46**	-15.51
			(15.93)	(25.09)
Observation Treatment*Female			41.55**	31.46
			(17.07)	(27.42)
Financial Treatment*Observation Treatment				-46.46
				(31.71)
Financial Treatment*Observation Treatment*Female				37.52
				(34.62)
Observations	5051	4300	2626	2326
Clusters	.	263	.	158
Dep. Var. Mean	365.4	365.4	365.4	365.4
Dep. Var. SD	164.7	164.7	164.7	164.7

Notes: This table presents the relationship between the employee's performance evaluation score, the treatment status and gender, controlling for employee effort.

- The dependent variable is the employee's evaluation score converted into the associated raise value in USD for that score.
- *Female* is a dummy for whether the employee is female. *Financial Treatment* is a dummy which is 1 if the teacher's school was assigned to have their evaluation determine their raise or 0 if their evaluation was just for feedback purposes. textitObservation Treatment is a dummy which is 1 if the teacher was randomly assigned be observed more frequently by their manager and 0 otherwise.
- Standard errors are clustered at the school level (the unit of randomization). All regressions control for teacher value-added, classroom observation score in each of the 12 dimensions of the CLASS rubric (Pianta et al., 2012), attendance, self-reported time use, intrinsic motivation and locus of control. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table 8: Manager Beliefs by Treatment

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

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

Table 9: Effect of Financial Stakes by Manager Type

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

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

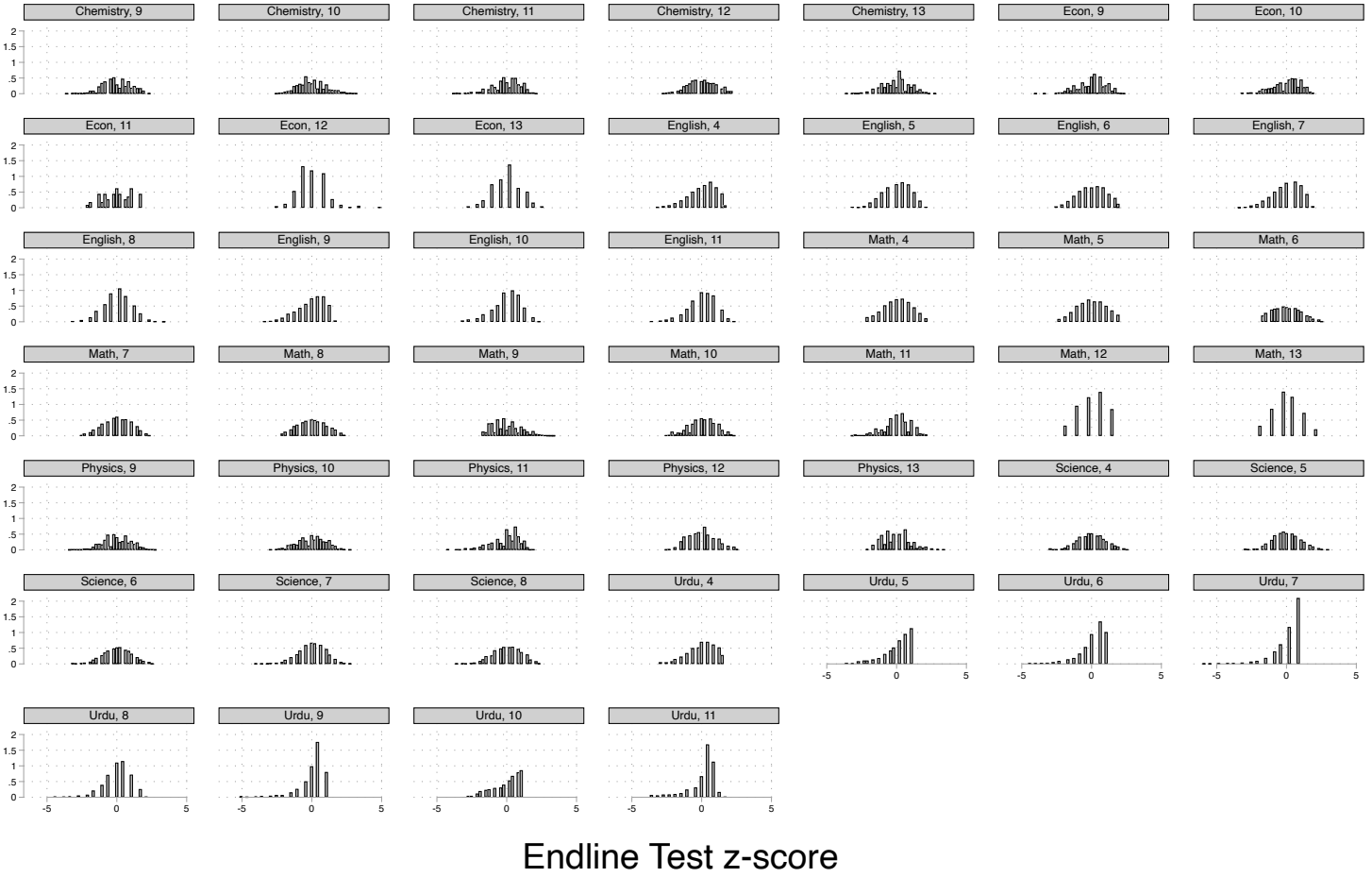
Table 10: Effect of Information by Manager Type

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

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

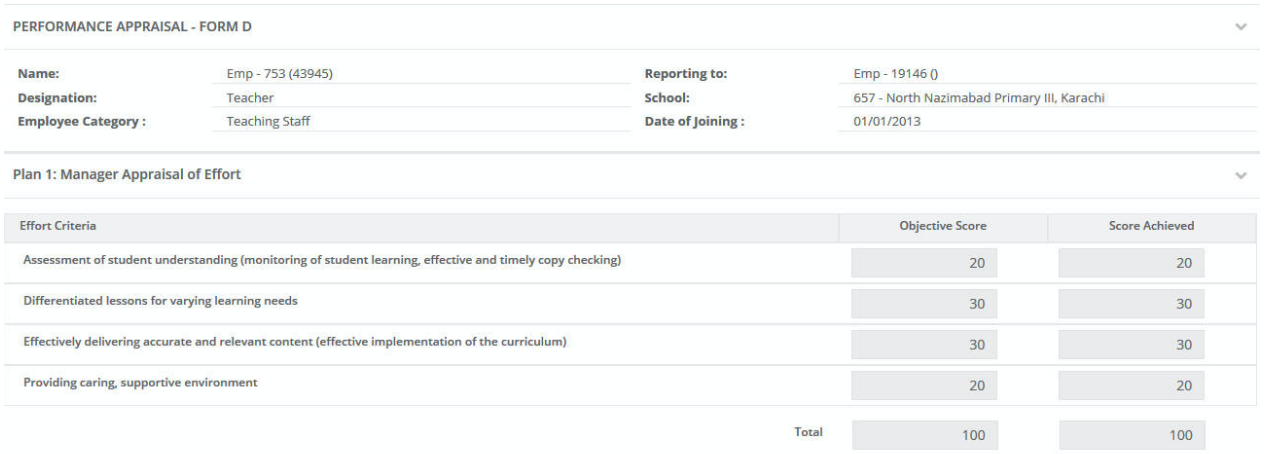
Appendix A - Supplementary Tables and Figures

Figure B1: Distribution of Endline Test Scores



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

Figure B2: 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.

Table B1: Baseline Covariates

Variable	(1) Control N/ [Clusters]	(1) Control Mean/ SE	(2) Objective Treatment N/ [Clusters]	(2) Objective Treatment Mean/ SE	(3) Subjective Treatment N/ [Clusters]	(3) Subjective Treatment Mean/ SE	T-test Difference (1)-(2)	T-test Difference (1)-(3)	T-test Difference (2)-(3)
Panel A: Teacher Characteristics									
Performance evaluation score	656 [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 B2: Raise Amount by Treatment and Gender (without Effort Controls)

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

Notes: This table presents the relationship between the employee's performance evaluation score, the treatment status and gender, but does not control for teacher effort.

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

Table B3: Teacher Characteristics - Survey Items

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

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

Appendix Appendix A.A VA Calculation

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

$$y_{ijkcst} = \beta_0 + \sum_s \beta_s y_{ijkcs,t\text{-}1} \mathbb{1}[\textit{subject-grade} = s] + \sum_s \alpha_s y_{ijkcs,t\text{-}2} \mathbb{1}[\textit{subject-grade} = s] \quad (15)$$
$$+ \sum_s \gamma_s \bar{y}_{\text{-}ijkcs,t\text{-}1} \mathbb{1}[\textit{subject-grade} = s] + \chi_{st} + \psi_k + v_{ijkcst}$$
$$\textit{where } v_{ijkcst} = \mu_j + \theta_{ct} + \epsilon_{ijkcst} \quad (16)$$

where y_{ijkcst} is the test score for child i with teacher j at school k in class c in subject-grade s in year t. We regress these test scores on the student's one-year, $y_{ijkcs,t\text{-}1}$, and two-year, $y_{ijkcs,t\text{-}2}$, lagged test score in the given subject and the class's average lagged test score, $\bar{y}_{\text{-}ijkcs,t\text{-}1}$. We allow the coefficients on lagged test scores (β_s, α_s and γ_s) to vary across subject-grade. χ_{st} captures subject-grade-year shocks. ψ_k captures school-specific shocks. The residual, v_{ijkcst}, is the combination of teacher effects μ_j, classroom effects, θ_{ct}, and student-time specific shocks, ϵ_{ijkcst}. To isolate the teacher component, we use the residuals, v_{ijkcst}, to construct an empirical Bayes estimate of teacher value-added. We compute the average weighted residual and shrink by the signal variance to total variance ratio (Kane and Staiger, 2008).[13] Teachers for which we have few student observations are shrunk toward the mean teacher value-added (normalized to be zero).[14]

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

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

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