Managerial Gender Bias in Performance Evaluation*

Christina Brown†

March 26, 2026

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 2,000 employees in 150 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 minimal difference in scores between men and women. This holds even when controlling for a rich set of controls of teacher productivity, such as value-added, clock in and out time, time use, and pedagogy measured via classroom observations. In contrast, when principals' evaluations determine teachers' end of year raise, we see that female teachers receive 20% lower raises, controlling for productivity. However, when principals are randomly assigned to conduct more frequent classroom observations of the teacher, this increases the evaluation of female teachers and closes more than half of the gender gap under financial stakes. To understand mechanisms, I conduct a follow up vignette survey to test whether our results are due to differential manager expectations about employee reactions to low raises (e.g. higher turnover by men) or differential perceived

*Funding for this project was provide by the UK Foreign, Commonwealth & Development Office, awarded through the J-PAL Post-Primary Education Initiative and RISE Programme, the Weiss Family Fund, CEGA, the Strandberg Fund, 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)

deservedness of scarce financial resources. The results favor this second explanation as managers favoring single-earner (lower income) households over dual-earner (higher income) households, which is highly correlated with employee gender. Combined this suggests that improving the accuracy of manager information could close the gender gap in performance evaluations, even in high stakes settings.

1 Introduction

The issue of gender discrimination in the workplace remains a significant global challenge, particularly in sectors where performance evaluations are closely linked to compensation and career progression. While gender biases in pay, promotions, and hiring have been widely studied, the underlying mechanisms driving these biases remain insufficiently understood. In many low-income countries, where female labor force participation is already lower than in higher-income countries, such biases can be even more pronounced. For instance, in Pakistan, women are more than three times less likely to be employed than men, with significant disparities even in female-dominated professions such as teaching, where women earn only 70 cents for every dollar earned by men. This paper seeks to contribute to this literature by exploring the extent to which financial stakes and manager discretion exacerbate gender biases in employee evaluations, using a large-scale field experiment in Pakistan's educational sector.

In this paper, I use a large-scale experiment with 200 managers and 2,000 teachers across 150 schools in Pakistan and vary features of the personnel system. This experiment allows me to answer two key research questions: How do personnel policies affect the extent of gender bias in employment outcomes? To what extent can we attribute this bias to statistical, taste-based or other forms of discrimination? The experiment allows me to test the conditions under which gender bias exists and better understand whether well-designed HR systems can constrain manager bias.

In order to better understand the mechanisms at play, I construct a conceptual framework which features a manager deciding how to allocate raises within a performance evaluation system. The manager observes a noisy signal of worker productivity, and then decides whether to use their own discretion in bumping the employee's raise up or down relative to observed productivity. In deciding whether to use discretion, managers trade-off between two competing forces that they expect to experience as a result of the selected raise: i) a cost from giving a low raise to an employee, which can vary across employee type and is increasing in what fraction of the raise is determined by the manager (versus other factors), and ii) a cost for giving inaccurate evaluations, which is decreasing in the noisiness of the production function. Disparate outcomes by gender arise through managers facing a larger cost of giving low raises to male employees, conditional on observed productivity.

To test this framework, I conducted a two-year experiment with a large network of schools in Pakistan to see whether the extent of gender bias in the evaluation of teachers varies by the financial stakes of the evaluation for teachers (i.e. does the manager's evaluation determine

the teacher's raise versus when the manager's evaluation is purely for feedback). Schools are randomized to either the "financial stakes" evaluation condition or the "no financial stakes" evaluation condition, where the managers evaluation has no financial stakes and the raise is determined by student test score performance or is flat across all teachers in the school. I also vary the noisiness of the production function by randomizing at the teacher level how frequently managers are required to conduct classroom observations of teachers. This increases the number of observations treated teachers receive relative to control by 50% in the three-month period before the evaluation score is submitted by managers.

To capture a comprehensive picture of teachers' actual productivity, I use rich data on teachers' value-added, daily attendance, time use, and measures of pedagogy from over 6,000 hours of classroom videos coded using the CLASS observation rubric. I also draw on detailed teacher, manager and student surveys to attempt to unpack the mechanisms at play.

Consistent with the framework, I find that when evaluation scores have no financial stakes for employees, male and female teachers received very similar scores, controlling for teacher productivity. However, when the evaluation score determines the employee's raise, men received a 21% higher raise than women.

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, they are better able predict teachers' actual performance on a number of dimensions. In turn, male and female teachers who were observed more frequently receive very similar evaluation scores. Whereas, for teachers who were observed less frequently, I find male teachers receive 22% higher raises than women 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 over a half.

We may be concerned that these results are driven by other changes in manager or teacher behavior in response to the treatments, such as changes to manager effort put into the evaluation process when there are financial stakes or differential teacher effort changes by gender. However, I do not find evidence in favor of these alternative mechanisms. Using metadata on the evaluation process and detailed data on teacher effort from classroom videos, I do not find evidence of differences by treatment, suggesting these are unlikely to be primary explanations for our findings.

The two primary remaining channels that could explain these findings are: i) managers anticipate differential responses to low raises by male versus female teachers (for example, higher likelihood to complain, reduced provision of school public goods, or higher turnover) and ii) differential deservedness of high raises by male teachers (for example, due to male

teachers being from primarily single-earner households versus female teachers being from primarily dual-earner households).

To test these two channels, I conduct a follow-up vignette survey with over 200 managers where I present hypothetical teacher profiles and ask managers to rank the quality of each teacher. The teacher profiles list a number of characteristics, including name, years of experience, test score growth, classroom management skills, and their spouse's employment status. Across survey subjects, I also vary whether the hypothetical ranking they perform is framed as selecting a performance evaluation score for the employee versus a performance evaluation raise. Finally, I present managers with a single teacher profile and ask managers to predict what they believe would happen if a given score/raise was allocated to that individual. Managers rate how likely that individual would be to look for another job, complain, etc.

Using the vignette survey, I find evidence in line with the latter channel (differential deservedness) but not the former channel (differential response to low raises). Managers are significantly more likely to give a high evaluation to profiles whose spouse is listed as not working when the decision is framed as a raise, but not when it is framed as a score. This is true for both male and female single-earner household profiles. However, I do not find a statistically significant difference for male versus female-named profiles in the likelihood a manager says there would be an adverse reaction to a low raise (such as reducing public good provision, complaining, or quitting). This result is consistent with findings from the administrative data that show when teachers receive a lower evaluation score for exogenous reasons, we do not see a differential likelihood in subsequent turnover by gender.

The paper makes three main contributions: First, I show that in this setting differences in the evaluation of male versus female employees by managers is not about differences in perceived competency or productivity but instead about the allocation of financial rewards to male employees. This complements other work on the extent of gender bias in different aspects of the employment process (Beg et al., 2021; Blau and Kahn, 2017; Grissom and Bartanen, 2022). The RCT design allows me to distinguish between the role of financial stakes for employees versus general bias in the evaluation of worker ability, and demonstrate how the former increases the extent of gender bias. This 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.

Second, I build on previous work that shows better information can lower the extent of

bias. Existing evidence, however, typically thinks of information as decreasing the extent of statistical discrimination, allowing managers to rely less on group attributes if they have better candidate specific information (Bohnet et al., 2012; Bartoš et al., 2016; Laouénan and Rathelot, 2022). This paper demonstrates a different channel through which information may lower bias, by making it less likely that managers will use discretion and exhibit financial favoritism toward male employees.

Finally, this paper builds on a burgeoning literature on the way employers may use hiring and wages to achieve social outcomes, such as redistribution (Macchi and Stalder, 2023; Swanson, 2024). I show that most of what looks like financial bias in favor of male teachers could actually be explained as financial favoritism toward lower income per capita households due to the high correlation between gender of the employee and whether the house is a single or dual earner household.

2 Gender, Teaching and Manager Bias

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

Studying the extent of gender bias in a setting with a large fraction of female employees and managers provides an interesting test because many of the typical explanations for gender gaps in wages are less likely to hold in this setting. In particular, we argue the typical channels of statistical and taste-based discrimination on ability/productivity are second order relative to other mechanisms in this setting.

2.1 Fact #1: Significant differences in teacher wages by gender

Across two different sample of teachers in Pakistan, we find a persistent gender wage gap. Women earn, on average, 20% less than men using data from 2500 private and public schools across Punjab province from Bau and Das (2020). While more narrow in sample, this data has the distinct advantage that we can control for school fixed effects, experience and credentials. In addition, the data has teacher value-added in each school subject and the

number of days in the last month the teacher was absent. These controls help to remove some concerns that differences in wages may be due to sorting into different schools or differential performance. Restricting to private schools, table A1 shows that the unconditional gender wage gap is 46%. When we control for value-added, attendance, credentials and school fixed effects, this gap shrinks a tad to 41%. As we might expect, the gap is much smaller in public schools (7% with all controls) which follow a standardized salary schedule with much less scope for discretion in wage setting.

In the RCT sample, there is an unconditional gap of 31% and are able to reduce it to 11% when we include rich measures of teaching quality (value-added, classroom observation score along 16 dimensions), experience, school, grade and subject fixed effects. This data is the most narrow in terms of representativeness, coming from a chain of 250 high-fee private schools spread across Pakistan, but we have the most granular performance measures helping to further alleviate other supply-side differences.

2.2 Fact #2: Minimal difference in employee ability by gender

To understand whether there is scope for statistical discrimination in our setting, I compare teacher characteristics along a number of dimensions. Figure 2 shows the distribution of teacher experience, attendance, value-added and classroom observation scores by teacher gender. For the first three characteristics, the distributions are nearly identical. Women however, have higher classroom observation scores.

We can also test whether there is a difference in the noisiness of teacher's performance by gender. To do this, we compare teacher value-added in the prior year versus the current year. Figure 3 presents a binned scatter plot of lagged value added versus current value-added is nearly identical for men and women.

2.3 Fact #3: Minimal gender bias on teacher ability

However, even if there are not true differences in ability by gender, managers may have incorrect beliefs about the true distribution. 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 1 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 A5 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. In contrast, we find a very strong relationship between the manager's rating and other performance characteristics like value-added and attendance.

One concern with using vignettes are that they simply do not have any predictive power. This could be the case if, for example, survey respondents 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 A6 shows the same specification as in table A5, 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. 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. Given these results, we therefore construct a framework which does not feature taste-based or statistical discrimination on teacher *ability/productivity* specifically but does result in disparate financial outcomes by gender.

3 Conceptual Framework - High Stakes Evaluations Under Uncertainty

Given the minimal scope for statistical or taste-based discrimination on employee ability, what could explain disparate wages by gender? In this section, we outline a model in which managers try their best to accurately evaluate their employees but have imperfect information about worker ability. When the evaluation is high stakes (has a financial consequence), managers derive higher disutility from giving raises to certain employees. This tension between accuracy and preferential treatment, results in managers giving certain employees the "benefit of the doubt" only when they have weak information about employee ability.

Consistent with the descriptive findings from the previous section, our model has no differences in the mean or variance of employee productivity by gender, and there is no taste-based gender discrimination by any managers on worker ability. 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 the typical discrimination channels do not hold.

The performance evaluation system takes place in three stages. First, employees work and produce some output which is observable to the manager and a function of true productivity. Next, managers select a raise for their employees which is a function of observed output and manager discretion. Finally, managers experience utility from the discretion they exerted in setting wages: i). they receive disutility from giving low raises (e.g. due to employee turnover, complaints, etc) and ii). they receive disutility from inaccurate raises (e.g. audits, psychic cost of lying, etc). Managers decide on the raise to give which minimizes these two disutilities.

Worker Output Employee i produces output y_i, which is the sum of their true productivity, 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.[1]

Manager Chooses Performance Raise Managers provide a performance raise for employees after observing y_i, and the raise is function of both this output and a discretionary component, d_i which is selected by the manager.

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

Managers ultimately select the value of this discretionary component to maximize their utility.

[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.

Consequence of Raise Decision for Managers There are two competing forces managers experience once the raise is set: i). disutility from giving low raises to certain individuals and ii). disutility from inaccurate evaluation of employee performance.

First, employees dislike receiving low raises, and will cause disutility to their manager as a result, which the extent of disutility decreasing in raise amount. This could be in the form of complaining to the manager, reduced cooperation, higher likelihood of turnover, etc. Alternatively, we could think of this as positive utility managers receive from giving higher raises to certain individuals. This could be due to differences in perceived deservedness, favoritism, bias or behaving consistent with social norms. In section 6, we will test possible sources of this disutility (employee backlash versus manager's internalizing worker utility). This expected "inconvenience" cost managers face is:

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

where c is the scale of cost (which depends on how much of the raise is determined by the manager versus other features) and ρ_i is the disutility for a given employee, conditional on the raise received. ρ_i can vary across employee, for example, it may cause a manager tremendous disutility to give a low raise to a colleague they have been friends with for decades or to a employee who is likely to quit if they receive a low raise.

Second, managers experience disutility from inaccurate evaluations. There could be several different channels micro-founding this disutility, such as the psychic cost managers face from lying, the long run incentive for managers to maintain the legitimacy of the performance evaluation system and the risk of punishment from higher ups within the organization if caught. This cost is a function of the difference between the raise given and the employee's true performance θ_i, since $r_i = \theta_i$ is the output-maximizing performance evaluation system.[2] The expectation of the inaccuracy cost at the time of selecting the raise is:

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

where p is the "price" of inaccuracy. We assume the cost takes this quadratic form in $r_i - \theta_i$ as it is hard to identify small discrepancies in scoring but easier to identify larger ones or blatant manager favoritism. This term could capture both the increased likelihood of audit and the larger punishment for large discrepancies. To determine p, we first solve

[2]We assume firms and managers are able to unearth θ through a costly information gathering process which they can do in the event of a complaint by an employee or a random audit.

for manager's optimal level of discretion for a given innaccuracy price p. Then we solve the firm's problem of optimally setting p.

Manager's Choice of Raise Managers select the discretionary component of the raise to minimize the inconvenience and inaccuracy costs they expect to face in the next period[3]:

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

Firm Setting Punishment "Price" The firm understands that managers do not have perfect information about worker ability at the time of setting raises, so they choose the "price" of inaccuracy to maximize the expected value of this punishment system (benefits - costs):

- Benefit of punishing (making discretion costly): $\mathbb{E}[-pd_i^2]$
- Cost of punishing (unnecessarily punishing managers for noise in the production function): $\mathbb{E}[p(2d_i\epsilon_i + \epsilon_i^2)]$

Firms select the value of p which maximizes the benefits minus the costs for the average level of c_i and ρ_i.

$$\max_p \mathbb{E}[-pd_i^2] - \mathbb{E}[p(2d_i\epsilon_i + \epsilon_i^2)] = \max_p -p\bar{d}^2 - p\sigma_\epsilon^2 = \max_p -p(\frac{\bar{c}\bar{\rho}}{2p})^2 - p\sigma_\epsilon^2$$

Taking the first order condition with respect to p, setting it equal to zero and replacing $(\frac{\bar{c}\bar{\rho}}{2})^2 = 1$ as it is a constant, we have:

$$\begin{aligned} &= (\frac{1}{p^2}) - \sigma_\epsilon^2 = 0 \\ p &= \frac{1}{\sigma_\epsilon} \end{aligned}$$

Therefore, the inaccuracy "price" inversely varies with the noisiness of the production function. This relationship would also be consistent if the inaccuracy cost is mostly an internal psychic

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

cost of lying that managers suffer. We would expect managers would "punish" themselves less if they are uncertain about the worker's true productivity and therefore do not actually know whether they are being inaccurate.

We can then substitute the expression for p into the optimal discretion function, so we have that $d_i^* = \frac{c\rho_i\sigma_\epsilon}{2}$. Therefore, an individual's raise will be $r_i^* = y_i + \frac{c\rho_i\sigma_\epsilon}{2}$.

Gender Gap in Raises Men and women's ability is draw from the same distribution $\Theta \sim \mathcal{N}(\mu, \sigma_\theta^2)$. The noisiness in the observation of output, ϵ is also drawn from the same distribution. We assume $\rho_m > \rho_f$. This difference could arise from differences in predicted response or perceived responsiveness. The difference in expected raises, conditional on ability, by gender then is:

$$\begin{aligned} \frac{\partial r_i^*}{\partial male}|_{\theta_i} &= y_m^* - y_f^* + \frac{c\sigma_\epsilon}{2}(\rho_m - \rho_f) \\ &= \frac{c\sigma_\epsilon}{2}(\rho_m - \rho_f) > 0 \end{aligned} \tag{7}$$

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

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

As we would expect, the raise is increasing in the scale of the inconvenience cost and noisiness of the production function. These parameters also affect the size of the gender gap: Effect of changes in: i). the dis-utility of low raises (c) and ii). the accuracy of information managers have (σ_ϵ)

1. Increased cost increases gender gap: $\frac{\partial^2 r_i^*}{\partial c \partial male}|_{\theta_i} = \frac{\sigma_\epsilon}{2}(\rho_m - \rho_f) > 0$
2. Increased noise increases gender gap: $\frac{\partial^2 r_i^*}{\partial \sigma_\epsilon \partial male}|_{\theta_i} = \frac{c}{2}(\rho_m - \rho_f) > 0$
3. Positive cross partial of cost and noise $\frac{\partial^3 r_i^*}{\partial c \partial \sigma_\epsilon \partial male}|_{\theta_i} = \frac{1}{2}(\rho_m - \rho_f) > 0$

4 Experimental Design

4.1 Mapping the Model to an Experiment

To test predictions 1-3, we need variation in what fraction of the raise is determined by the manager versus something else (c) and the noisiness between what managers observe and teacher's actual ability (σ_ϵ). 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 4.2.

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 4.4, we discuss each measure.

Finally, in section 6, we attempt to unpack the sources of the inconvenience costs, C_i. We test whether these costs mostly stem from managers expecting a differential negative response from male versus female employees (e.g. higher rates of turnover or reduced contribution toward school public goods) or whether there is a difference in perceived deservedness of higher salaries for men versus women.

4.2 HR Policy Interventions

In order to test our hypotheses, we introduce two variations in how managers within the school system evaluate their employees. The randomized controlled trial was conducted from October 2017 to June 2019 with a private school chain that operates nearly 300 schools located across Pakistan. Figure A1 presents the timeline of interventions and data collection activities.

Performance Evaluation Cycle In all schools, employees receive an annual performance review by their manager. 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 spot checked 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.[4]

- **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.[5,6] The contract applied to all core teachers (those teaching Math, Science, English, Urdu, and Social Studies) in grades 4-13. Elective teachers and those teaching younger grades received the status quo contract. All three contracts have equivalent budgetary implications for the school. I over-sampled the number of subjective treatment arm schools due to partner requests, so the ratio of schools is 4:1:1 for subjective treatment, objective treatment, and control, respectively.

Treatment 2: Increased Observation of Employee Effort In addition to the variation in financial stakes of the performance evaluation, I vary how often managers conduct classroom observations of certain teachers. At the beginning of the second semester of the intervention year, all managers receive a training from the school system on how to use a new classroom observation tool to record their notes and feedback during classroom observations. They are

[4]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.

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

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

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.

4.3 Data

I draw on four main sources of data from the RCT: i). school system administrative data, ii). student test scores, iii). manager and employee surveys, and iv). classroom observations for the main results.

Type	N	Source	Outcomes
Teachers			
"Ground truth"	1,500	Class video	Rubric covering 20 aspects of pedagogy (Araujo et al, 2016 + others
	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 and TIMSS) and a locally used standardized test (LEAPS) were also included to benchmark student performance.[7]

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 A8 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.[8] 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

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

[8]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.

dimensions.[9,10] We also recorded whether teachers conducted any sort of test preparation activity and the language fluency of teachers and students.

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

To test mechanisms and understand external validity, I use four additional sources of data: i). a followup vignette survey with school principals (described in section 6.1) ii). the 2018 Pakistani Social and Living Standard Measurement survey (PSLM), iii). data on teacher quality and salaries from Bau and Das (2020) and iv). the 1980-2021 rounds of the Panel Study of Income Dynamics (PSID).

[9]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).

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

4.4 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 A2 shows the distribution of teacher value-added for the 3,687 teachers who teach in the school system at baseline.

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.

4.5 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. Annual school fees are $900 USD. Table 1, panel A, presents summary statistics for our sample teachers compared to a representative sample in the US (National Center for Education Statistics, 2011). Our sample is mostly female (81%), young (35 years on average), and have worked in the school system for a mean of 5 years. Nearly all teachers have a BA, and 68% have some post-BA credential or degree.

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 1, panel B and C presents information about managerial duties compared to a US sample of principals. Like in the US, our managers are generally older (45 years old) and less likely to be female (61%) 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 A7 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.

5 Results

In the following section, we will look at the effect of the 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 first show evidence that the intervention was implemented effectively and well understood by teachers and managers. Then 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).

5.1 Effect of Financial Stakes on Evaluation Scores

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 A6 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.[11]

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:

$$PredictedRaise_{is} = \beta_0 + \beta_1 FinancialStakesTreatment_s + \beta_2 Female_i + \beta_3 FinancialStakesTreatment_s * Female_i + \chi_i + \gamma_s + \epsilon_{is} \quad (9)$$

where the dependent variable is the teachers' predicted end of year raise using their manager-assigned performance evaluation score (raise percentage for evaluation score group*base salary). 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. $FinancialStakesTreatment_s$ 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. The main figures and tables use double lasso to select among this set of controls and in the appendix, I show robustness to keeping all the controls in as well as removing all of them. γ_s are randomization strata fixed effects. 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).

Consistent with prediction 1, I find that managers allocate performance evaluation scores differently when there are financial stakes versus not and this particularly hurts female teachers. In figure 6, we see that men and women receive nearly identical predicted raises when there are no financial stakes. In the figure, each bar plots the average raise the teacher would receive if their raise was based on their performance evaluation score, by treatment status and gender, controlling for teacher productivity and restricting to the control group for the observation treatment. The second two bars show men and women would receive similar raises in the schools where there was no actual financial stake of the evaluation score,

[11] Discussion of these behavioral effects by teachers to the subjective versus objective treatment versus no performance pay is the focus of a companion paper (Andrabi and Brown, 2020)

with women earning just slightly more ($1306 for men versus $1315 for women). However, when there is a financial stake of the evaluation score, women receive significantly lower raises ($1446 for men versus $1101 for women), equivalent to 24% reduction compared to men.

Table 2, col 2 presents the results of eq. 9. As we saw in figure 6, the interaction between the financial treatment and gender (β_3) is large and highly significant, -$281 ($p < 0.001$). The coefficient remains large and significant when we include all the productivity controls (-$233, $p = 0.001$) or do not include any (-$291, $p < 0.001$) shown in tables 3 and4, respectively.

Interestingly, the coefficient is still negative but no longer significant when we adjust our outcome to be the predicted raise percentage rather than the raise amount (which is raise percent*base salary), shown in table A9. Since managers have a fixed quota of how many employees they can allocate to each evaluation score category, we can also test the effects using a rank-ordered logit (table A10), and results are similar. This gives us further insight into how managers are allocating resources. When there are financial stakes of their evaluation, managers give higher scores specifically to men with higher baseline salaries (not just to all men).

5.2 Effect of Information on Evaluation Scores

For the manager observation treatment, we first show that having managers do more classroom observations does, in fact, improve the information they have about teacher quality. Table A11 and figure A7 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. 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. 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 the classroom observation treatment versus those in the

status quo. Our main specification is:

$$PredictedRaise_{is} = \beta_0 + \beta_1 ObservationTreatment_i + \beta_2 Female_i + \beta_3 ObservationTreatment_i * Female_i + \chi_i + \gamma_s + \epsilon_{is} \quad (10)$$

where the dependent variable is the teachers' predicted end of year raise using their manager-assigned performance evaluation score. $ObservationTreatment_i$ is a dummy for whether the employee was assigned to be observed more frequently by their manager. $Female_i$, χ_i and γ_s are the same as in equation 9. 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 school level (the unit of randomization).[12]

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 closes substantially. In figure 7, each bar plots the average predicted raise based on evaluation score, by observation treatment status and gender, controlling for teacher productivity (restricting to schools with financial stakes). The first two bars show men receive significantly higher scores under the status quo level of observations, with an average difference of \$343 (22% of their raise, $p = 0.001$). The next two bars show that under the additional monitoring by managers, male and female teachers now receive much more similar raises, with an average difference of \$98. Table 2, col. 3, shows the interaction between the observation treatment and gender pooling schools with and without financial stakes ($\beta_3 = \$181$, $p = 0.027$). The results are again similar, if we include all the teacher productivity controls (\$207, $p = 0.01$) or none (\$161, $p = 0.052$), and are shown in tables 3 and 4, respectively.

5.3 Interaction of Financial Stakes and Information

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

[12]The observation treatment was first randomized at the school level (certain schools were selected to participate versus not) and then within selected schools, certain teachers were randomly selected to be observed. Because the financial stakes treatment is at the school level, for comparability, I cluster at the school level in all specifications.

stakes. The specification I test is:

$$\begin{aligned} PredictedRaise_{is} = \beta_0 &+ \beta_1 FinancialStakes_s + \beta_2 Female_i + \beta_3 FinancialStakes_s * Female_i \\ &+ \beta_4 Observation_i + \beta_5 Observation_i * Female_i \\ &+ \beta_6 FinancialStakes_s * Observation_i \\ &+ \beta_7 FinancialStakes_s * Observation_i * Female_i + \chi_i + +\gamma_s + \epsilon_{is} \quad (11) \end{aligned}$$

where the dependent variable, treatments and controls are the same as in eq. 9 and 10. β_3 tells us the effect of the finance treatment on gender bias for the observation status quo teachers. β_5 tells us the effect of the observation treatment on gender bias in the finance control schools. β_7 tells us whether there is a compound effect of the two treatments.

I find some weak evidence that the observation treatment has a larger effect under financial stakes, possibly because under no financial stakes men and women are already scored similarly. Table 2, col. 4 show the interaction of the two treatments. Under the observation status quo, the interaction of *Financial stakes*Female* is -$363. Whereas, it is -$200 under the observation treatment. However, the triple interaction coefficient β_7, while large in magnitude, is not statistically significant.

5.4 Heterogeneity by Manager Characteristic

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 across setting. To test this, we look at heterogeneous treatment effects by a variety of manager characteristics: gender, age, 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 5 and table 7 present results from eq. 9 and eq. 10 adding in an interaction with the respective manager covariate. One suggestive pattern we see is that the financial stakes actually have *less* of a negative effect on female ratings for managers who are more "biased" as measured from our survey though the difference is not statistically significant. 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. There is not a statistically significantly different response to the financial stakes or observation treatment by manager gender or age.

5.5 Alternative Explanations

There are several alternative ways that financial stakes may have changed the evaluation process differentially for male versus female teachers which has nothing to do with gender bias or two main mechanisms we will focus on in section 6. The primary alternative channels are when there are financial stakes: a). managers take the evaluation process more seriously/carefully (and are possibly more accurate), b). managers value different aspects of teacher effort and c). female versus male teachers differentially respond to the financial incentives along dimensions of effort we are not controlling for.

First, to test whether managers take the evaluation process more seriously when there are financial stakes, I compare the length of each criteria in terms of number of characters, how likely managers are to use round numbers in scoring each sub criteria (scores ending in 0 or 5), and the cross-criteria variance in scoring within each teacher. There is no statistically significant difference along any of these dimensions by whether the evaluation had financial stakes (table A14).

Second, to test if perhaps managers valued certain typed of teacher actions, I have a team classify each criteria into one of 27 areas of teacher performance (e.g. parent engagement, assisting with extracurriculars, providing feedback to students, etc). I then test along each of the 27 areas whether the criteria used to evaluate teachers is more likely to fall in that area when there are financial stakes. In total, I run 54 regressions (each area on the intensive and extensive margins) and plot the p-values (figure A8). The red line represents the distribution of p-values we would expect to see were we to conduct infinite tests for a true treatment effect of zero. Our distribution of p-values looks similar to what we would expect to see with no true difference by treatment.

Finally, in general, female teachers appear to be slightly more responsive to the financial stakes than male teachers on a subset of outcomes we can measure. Using midline classroom

observation data and endline teacher surveys and student tests, we can compare the effect of financial stakes of the performance evaluation score versus no stakes (pooling schools with a flat raise and test-score based performance raise) on 29 different measures of teaching quality (figure A9). Of those, I find no statistically significant difference by gender on 20 of the 29 measures. On eight measures, I find female teachers have a larger positive treatment effect (value-added, observation scores on positive classroom climate, teacher sensitivity, regard for student perspectives, diverse learning formats, student engagement, classroom environment and self-reported time spent lesson planning), and on one there are larger treatment effects for men (days present at school). For the observation treatment, 26 of the 29 measures do not have a statistically significantly different treatment effect for men versus women (figure A10). Three have larger treatment effects for male teachers: time spent grading and interacting with peer teachers (significant at the 10% level) and time spent interacting with managers (significant at the 5% level).

In order for this differential response to incentive pay by gender to inflate our results, it would have to be that there are other aspects of teacher quality that we have not measured (or have measured noisily) for which male teachers are *more* responsive to financial stakes. Since, if anything, women appear to be slightly more responsive in the aspects of quality we can measure, this confound seems unlikely. Given our results also appear very similar whether we include all productivity controls, double lasso selected productivity controls or no productivity controls, this does not appear to be major concern. Lastly, if it were simply that male teachers are just much better performers under financial stakes (in ways we can't control for with our measures), we would not expect for the observation treatment to have the effect it does in reducing the extent of financial favoritism toward male teachers.

6 Mechanisms

6.1 Vignette survey to test mechanisms

There are two remaining broad explanations which could explain the patterns that we observe: i). managers expect a different response to a low raise from men versus women (such as higher turnover or reduced effort) or ii). managers believe men are more deserving of higher raises (for example, because of differences in total household income). To test each of these possible explanations, I conducted a vignette survey with a separate sample of 240 public school principals who do not currently have a high stakes performance rating system of teachers in their schools. The vignette survey allows me observe how managers

respond to certain worker characteristics, holding other attributes fixed. Across managers I also vary features of the hypothetical evaluation decision. The survey was framed as trying to understand what managers value in teachers and how they would evaluate (or reward) different teacher attributes.

The vignette survey proceeds in three steps. First, managers are told they will be evaluating example teachers with characteristics similar to those of teachers in their school system. They are also told the features of the hypothetical evaluation system, such as whether the evaluation score would affect the teacher's pay. Next, managers are shown three profiles of teachers, describing the teacher's performance, experience and several demographic features, and managers are asked to rank the three profiles. Figure A11 shows an example of the three teacher profiles managers would see. Finally, managers are shown a single teacher profile and asked how likely different outcomes (such as whether the teacher would complain, look for a new job, etc) would be if the teacher was given a certain rank. Table A15 describes the full set of questions asked in the survey.

The features of the hypothetical evaluation system are randomized across subjects and are: - Financial stakes:

- Affects pay: "The score would affect the teacher's pay for the next year"

- Doesn't affect pay: "The score will **not** affect the teacher's salary or promotion opportunities"

The features of the teacher profiles that are shown are randomized within-subject and include the teacher's subject and grade taught, student test score growth, classroom management skills, attendance, accuracy of information, expected turnover, years working together and spouse's job. Table A16 describes the full set of feature attributes and possible values.

6.1.1 Differential response to wages

To test whether managers predict a differential response to low raises by gender, I show managers an example teacher profile and ask them what they expect would happen if that teacher were to receive a low raise. Because the profile's gender is randomly assigned and I can control for the other characteristics of the profile, isolating whether managers perceive men versus women will respond differently.

Across several analyses, I do not find evidence for differential response to low raises being an important channel. Overall, I do not find a statistically significant difference by profile gender in the likelihood that managers believe a teacher would complain, job hunt, suffer financially, or contribute less to school public goods when they receive a low raise. There

is also not a statistically significant difference by profile gender in how likely managers are to say they would feel bad for the teacher or that they would feel good about the decision made. Figure 8 present managers' average responses by profile gender.

In addition, when I look at historical administrative data for the school system over the last 5 years, I do not see that men versus women are more likely to turn over after receiving a low raise. However, this evidence is just suggestive as there may be important other confounds, such as differences in teacher ability for those receiving a low raise by gender, which may mask this mechanism when we just look at turnover in the cross-section.

Second, at the end of the survey I ask managers directly whether men versus women, younger versus older, and inexperienced versus experienced teachers are more likely to complain be upset when receiving a low raise. Half of managers say men and women would be equally upset and likely to complain. Another quarter say men are more likely and the final quarter say women are more likely. Figure A12, panel A shows the distribution of manager responses. Importantly these results are not due to managers just being unwilling to say a certain group would respond differentially. Half of managers state that experienced teachers would be more or much more likely to complain and be upset compared to novice teachers (figure A12, panel C).

6.1.2 Differential deservedness of wages

To test whether managers exhibit differential deservedness of higher wages by gender, I vary the employment characteristics of the spouse on the teacher's profile, stating that the teacher's spouse either does not work, is a teacher, or is a doctor. For managers who are told the evaluation score would not affect pay, the spouse's characteristics do not affect the evaluation score, which is reassuring. However, when managers who are told their evaluation score would affect pay, teachers whose spouse does not work receive 0.20SD higher evaluation scores compared to those whose spouse works. Figure 9 shows the average evaluation score by financial stakes of the evaluation and spouse employment status. This difference in scoring exists for both male and female teachers. This suggests that managers are not explicitly prioritizing a certain gender but instead considering total household income in their evaluation decisions.

Could this channel explain the results of our main experiment? Yes, there is, in fact, a very strong correlation between gender and total household income in Pakistan, and so what appears to be bias against female employees could instead be favoritism toward lower-income households. Using data from the Pakistan Social and Living Standards Measurement survey

(PSLM) and restricting to primary and secondary school teacher, we can see that 42% of female teachers are married compared to 82% of male teachers. Among married teachers, 74% of female teachers' spouses work as compared to 20% of male teachers' spouses. Female teacher's total household income and total income per capita are 29% and 48% higher than male teachers, respectively (figure A13).

While I do not have total household income or spouse's employment status for our main RCT sample, I do have teacher age, and I find the results are stronger amongst teachers who are 30 years and older (figure A15). This would be consistent with a story in which managers show favoritism toward single-earner individuals with families.

7 Conclusion

In this paper, we explored how gender biases manifest in the performance evaluation of teachers in a large-scale experiment conducted in Pakistan. By manipulating the financial stakes tied to evaluation scores and increasing the frequency of manager observations, we were able to analyze whether male and female teachers are rated differently based on their productivity and effort. The experiment also tested how improved information through more frequent observations impacts the gender gap in raises when performance evaluations are tied to financial outcomes. The findings provide a comprehensive look at how subjective biases can shape employment outcomes even in the absence of overt discrimination.

Our analysis demonstrated that in schools where performance evaluations determined teachers' raises, male teachers received significantly higher raises compared to female teachers, even when controlling for productivity. This disparity largely disappeared when managers conducted more frequent classroom observations. This suggests that biases may be partially driven by a lack of detailed information about employee performance, with more informed managers less likely to favor male employees. Furthermore, the study found that the gender gap in raises was not as pronounced when evaluation scores had no financial consequences, indicating that the stakes associated with the evaluation process exacerbate the bias.

Finally, we unpack the mechanisms underlying the extent of manager bias and find evidence in support of managers showing favoritism toward lower income, single-earner households. We find less evidence that managers are giving male teachers higher raises because they anticipate more negative backlash from male teachers. This suggests that women who enter the labor force may be "doubly" punished – receiving lower wages and also having their spouse's incomes reduced. It also presents interesting areas for future work to

better understand how managers trade off profit maximization in their personnel decisions with preferences for redistribution.

8 References

References

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

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.

Bartoš, Vojtěch, Michal Bauer, Julie Chytilová, and Filip Matějka, "Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition," 2016, *106* (6), 1437–1475.

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.

Bohnet, Iris, Max H. Bazerman, and Alexandra Van Geen, "When Performance Trumps Gender Bias: Joint Versus Separate Evaluation," 2012.

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.

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.

Laouénan, Morgane and Roland Rathelot, "Can Information Reduce Ethnic Discrimination? Evidence from Airbnb," 2022, *14* (1), 107–132.

Macchi, Elisa and Jeremia Stalder, "Preferences for redistribution via employment and the organization of production in urban Uganda," 2023. Publisher: Open Science Framework.

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.

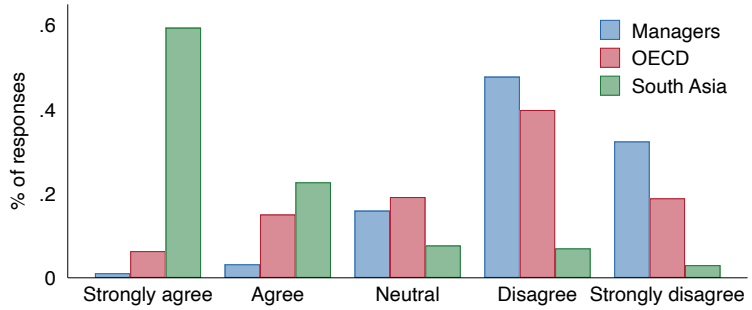
Pianta, Robert C, Bridget K Hamre, and Susan Mintz, *Classroom assessment scoring system: Secondary manual*, Teachstone, 2012.

Swanson, Nicholas, "All in the Family: Kinship Pressure and Firm-Worker Matching," 2024.

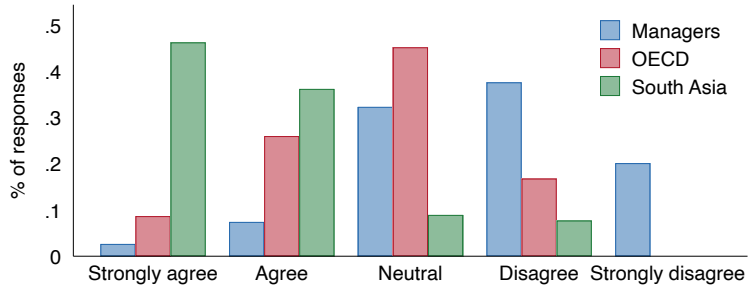
9 Figures

Figure 1: Managers' Stated Beliefs about Women and Employment

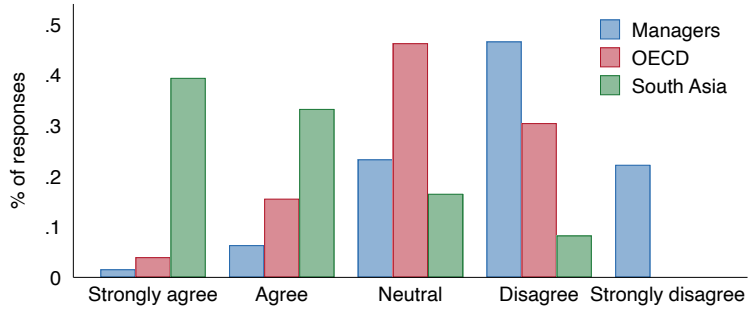
Panel A: When jobs are scare, men should have more rights to a job than women



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



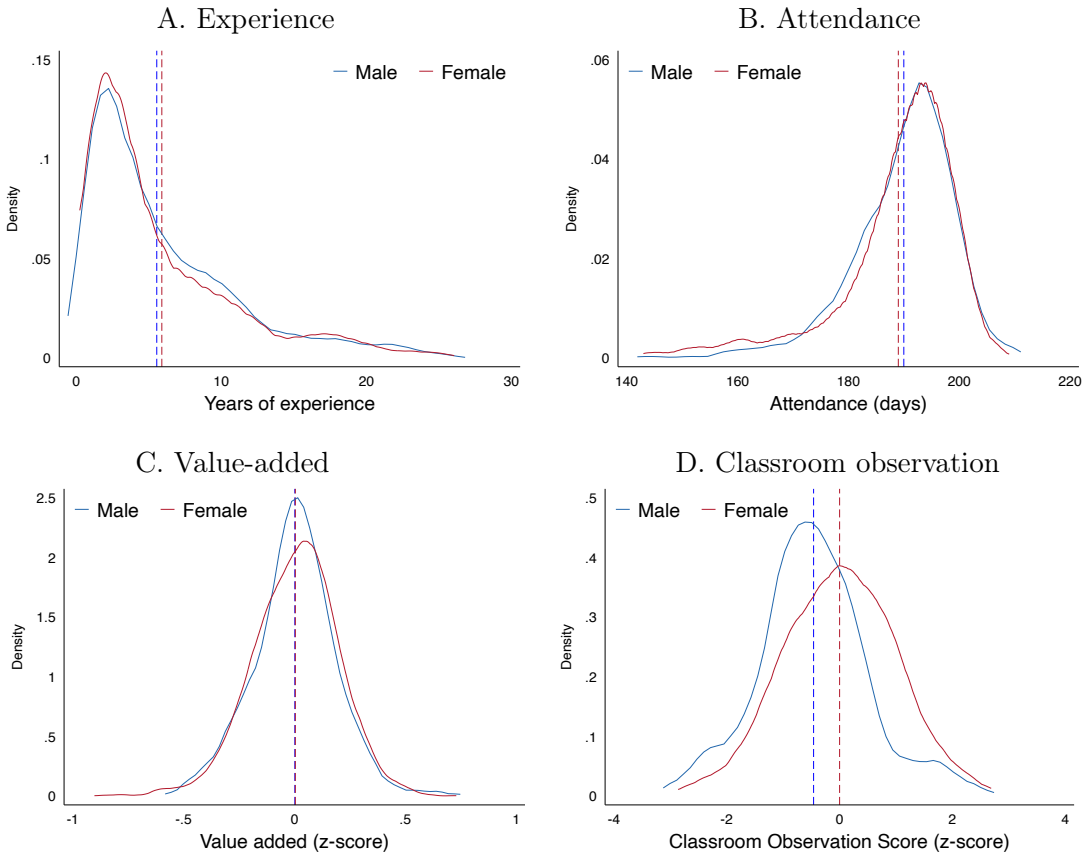
*Panel C: Men are better suited than women to teach math and science***



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"

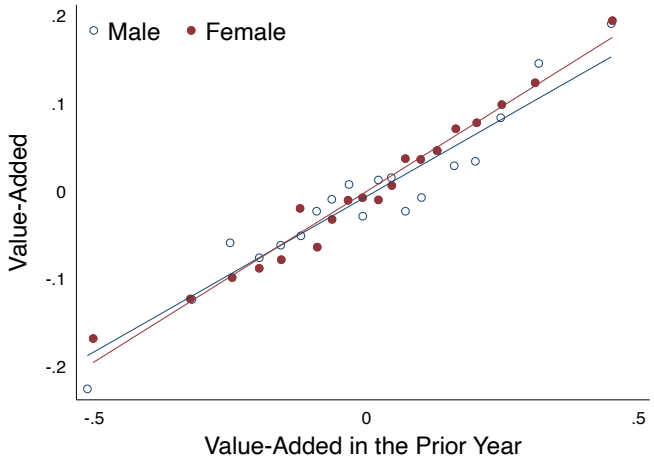
Figure 2: Teacher Effort and Productivity Measures by Gender



Notes: This figure shows the density plots of various measures of teacher effort and productivity by gender.

- Panel A shows the distribution of teacher years of experience by gender. Data comes from administrative employment records provided by the school system.
- Panel B shows the distribution of teacher days of attendance by gender. Data comes from administrative biometric clock in and out records provided by the school system.
- Panel C shows the distribution of teacher value-added by gender. Data comes from administrative test score data and class assignment records provided by the school system. A full description of the value-added calculation is provided in appendix Appendix A.A.
- Panel D shows the distribution of classroom observation scores by gender. Data comes from video tapes recorded and scored by the research team of teachers' lessons.

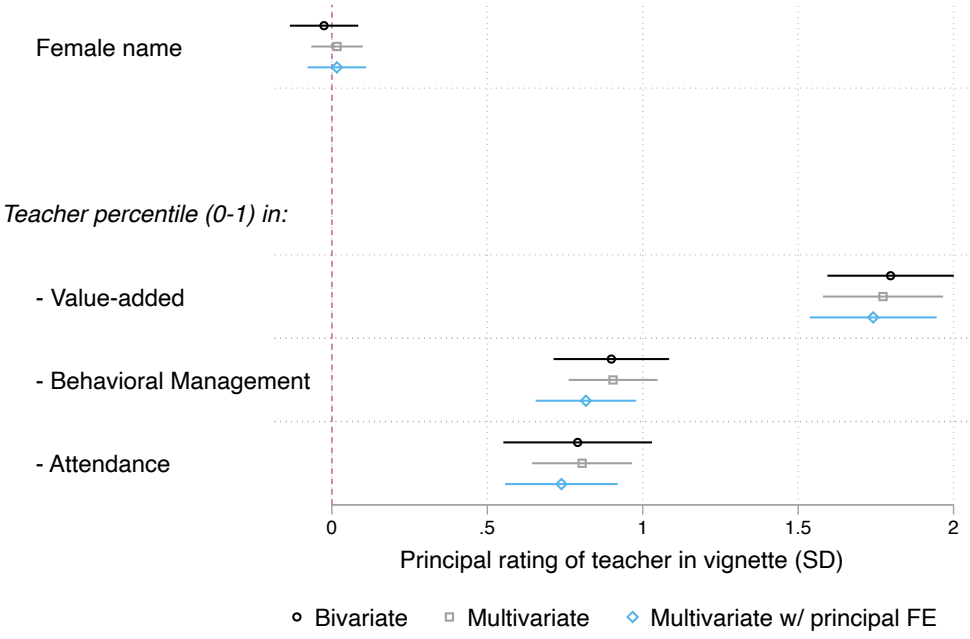
Figure 3: Year-to-year Correlation in Value-Added by Gender



Notes: This figure shows the year to year correlation of teacher value-added by teacher gender.

- The x-axis displays the teacher's value-added in year t, the y-axis displays the teacher's value added in year t+1.
- The solid, red points display ventile means for female teachers, and the hollow, blue points display ventile means for male teachers. The solid red and blue line show the linear best fit line for female and male teachers, respectively.
- Data comes from administrative test score data and class assignment records provided by the school system. A full description of the value-added calculation is provided in appendix Appendix A.A. The sample is necessarily restricted to teachers for whom we have at least two consecutive years of value-added measures.

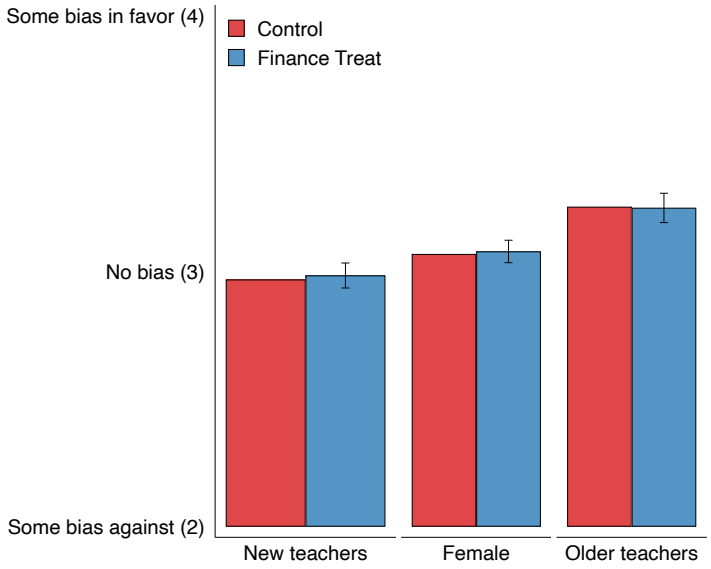
Figure 4: Manager Evaluation Score of Example Vignettes



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.

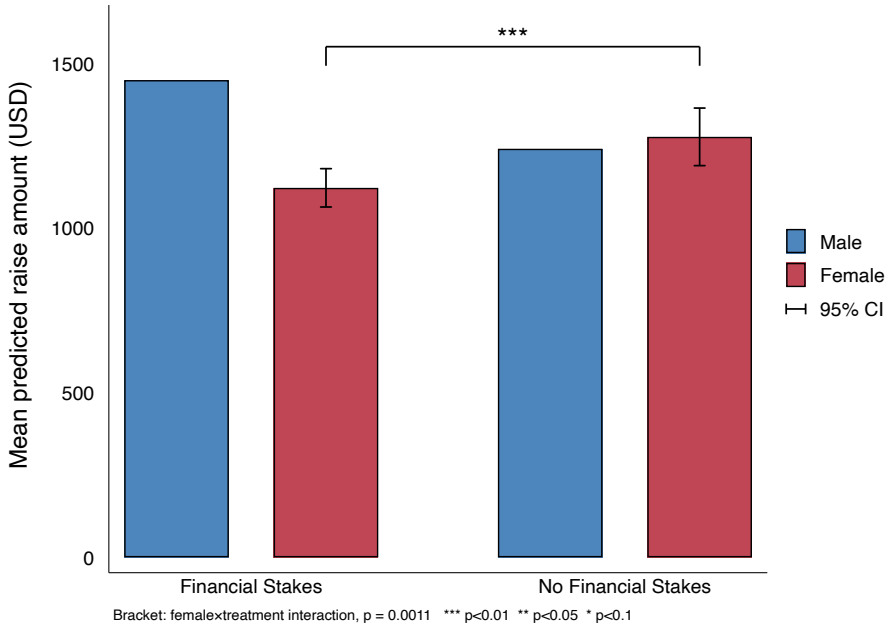
Figure 5: Teachers' Perceptions of Bias in Evaluation Scores by Treatment



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.

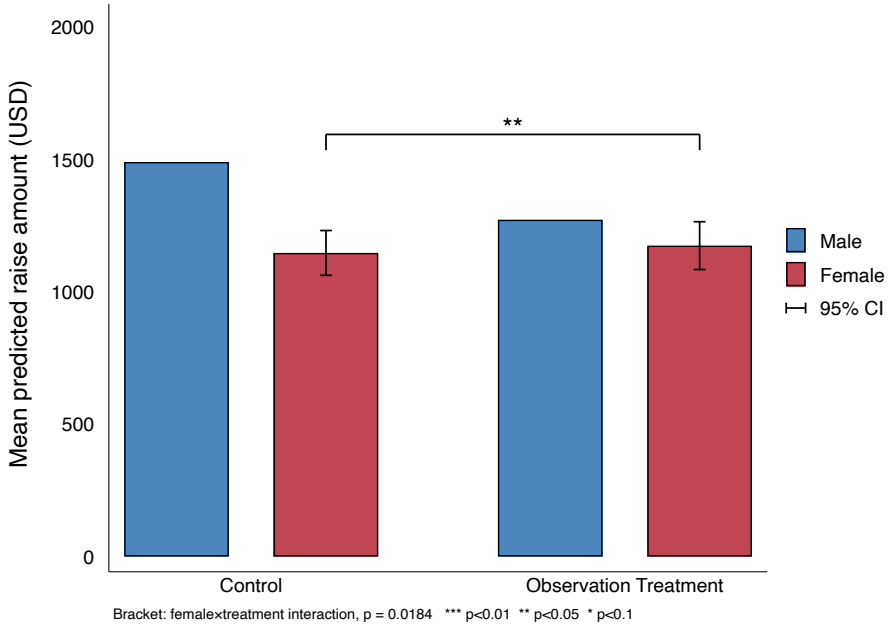
Figure 6: Effect of Financial Stakes Treatment on Gender Bias



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 financial stakes (observation treatment control group).

- 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). Figures control for double LASSO selected controls from the set of: 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. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

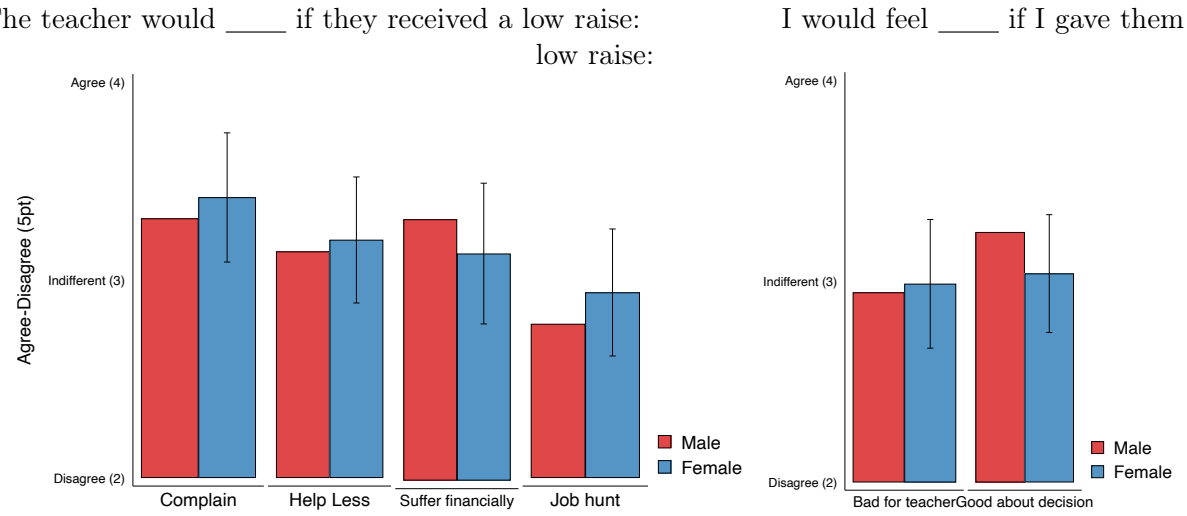
Figure 7: Effect of Employee Observations on Gender Bias



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 observation treatment status (financial stakes school sample).

- 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 school level. Figures control for double LASSO selected controls from the set of: 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. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

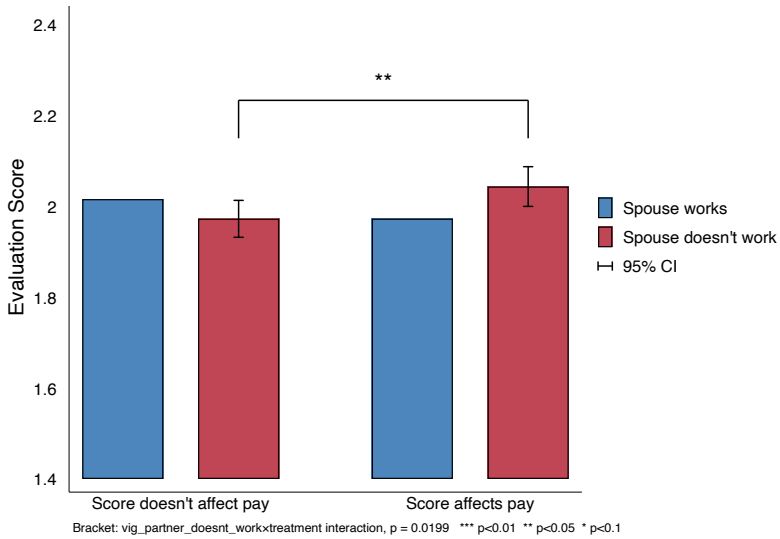
Figure 8: Managers' Predictions of Teacher Response to a Low Raise by Gender



Notes: This figure shows the mean manager response in the vignette survey sample to the predicted response to a low raise by vignette name gender.

- The y axis is mean Agree-Disagree response from Strongly agree (5) to strongly disagree (1) given by survey respondents after being shown a profile of an example teacher and asked a series of questions about the teachers response if they were to be awarded a low raise. The bars in red and blue show the average response when the vignette used a male versus female name, respectively. 95% confidence intervals are shown testing the equality responses when the vignette has a male versus female name.
- Data is from the manager vignette survey conducted with 239 principals in spring 2023.

Figure 9: Effect of financial stakes on evaluation score by spouse's employment



Notes: This figure shows the effect of financial stakes on the evaluation score by spouse's employment from the hypothetical vignette survey conducted with principals.

- The y axis is the average evaluation score given by survey respondents. The first two bars show the average rating by spouse's employment for respondents who were told the score would not affect pay. The second two bars show the average rating by spouse's employment for respondents who were told the score would affect pay. The bracket above the bars shows the statistical significance of the interaction between financial stakes and spousal employment.
- Data is from the manager vignette survey conducted with 239 principals in spring 2023.
- Standard errors are clustered at the manager level (the unit of randomization). $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

10 Tables

Table 1: Descriptive Statistics about Study Sample and Comparison Sample

	Study Sample		US Sample	
	Mean	St. Dev.	Mean	St. Dev.
Panel A. Teacher Characteristics				
Age	35.0	8.9	41.8	7.5
Female	0.80	0.40	0.76	0.43
Years of experience	5.1	5.2	13.8	9.6
Has Post BA Education	0.68	0.47	0.54	0.50
Salary, USD	4,000	1,700	52,400	18,400
Panel B. 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
Panel C. 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

Notes: This table reports summary statistics on teacher and principal characteristics for our study sample, and a comparison sample in the US. Data in panel A and B, 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. Col. 3 and 4, School and Staff Survey (National Center for Education Statistics).

Table 2: Raise Amount by Treatment and Gender, PDS Lasso Controls

	Predicted Raise Amount (USD)			
	(1)	(2)	(3)	(4)
Female	-106.8*	69.45	-194.9**	36.18
	(55.70)	(44.99)	(80.88)	(55.52)
Financial stakes		107.2		208.5*
		(82.99)		(112.4)
Financial stakes*Female		-280.8***		-363.3***
		(77.31)		(109.3)
Observation Treatment			-138.8*	1.142
			(78.86)	(95.34)
Observation Treatment*Female			180.5**	76.18
			(81.58)	(98.22)
Financial stakes*Observation Treatment				-213.7
				(134.4)
Financial stakes*Observation Treatment*Female				162.6
				(140.6)
Observations	1961	1961	1961	1961
Clusters	152	152	152	152
Dep. Var. Control Mean	1532.8	1532.8	1532.8	1532.8
Dep. Var. Control SD	830.2	830.2	830.2	830.2

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. *Observation 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 strata and double lasso selected controls from the set of: 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 3: Raise Amount by Treatment and Gender, All Controls

	Predicted Raise Amount (USD)			
	(1)	(2)	(3)	(4)
Female	-59.02	85.70*	-159.5**	36.24
	(41.39)	(44.11)	(66.61)	(60.48)
Financial stakes		57.82		155.2
		(73.74)		(100.5)
Financial stakes*Female		-233.2***		-311.3***
		(67.06)		(97.59)
Observation Treatment			-161.7**	-25.59
			(77.88)	(89.81)
Observation Treatment*Female			206.6**	106.0
			(80.03)	(91.12)
Financial stakes*Observation Treatment				-209.8*
				(121.9)
Financial stakes*Observation Treatment*Female				157.2
				(126.8)
Observations	1961	1961	1961	1961
Clusters	152	152	152	152
Dep. Var. Mean	1532.8	1532.8	1532.8	1532.8
Dep. Var. SD	830.2	830.2	830.2	830.2

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. *Observation 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 strata and 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 4: Raise Amount by Treatment and Gender, Just Strata Controls

	Predicted Raise Amount (USD)			
	(1)	(2)	(3)	(4)
Female	-120.3**	60.64	-198.2**	40.36
	(55.78)	(42.82)	(78.89)	(52.24)
Financial stakes		88.04		203.3*
		(84.36)		(108.9)
Financial stakes*Female		-291.2***		-377.4***
		(78.60)		(108.1)
Observation Treatment			-113.1	41.74
			(78.08)	(98.89)
Observation Treatment*Female			161.0*	57.11
			(82.04)	(99.83)
Financial stakes*Observation Treatment				-238.7*
				(138.4)
Financial stakes*Observation Treatment*Female				163.5
				(144.9)
Observations	1961	1961	1961	1961
Clusters	152	152	152	152
Dep. Var. Mean	1532.8	1532.8	1532.8	1532.8
Dep. Var. SD	830.2	830.2	830.2	830.2

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. *Observation 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 strata fixed effects. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table 5: Effect of Financial Stakes by Manager Type

	Predicted Raise Amount (USD)					
	(1) Male Manager	(2) Manager Age	(3) Avg. Bias	(4) Math Bias	(5) Jobs Bias	(6) Family Bias
Female	47.84	419.6**	-245.1	4.663	-73.72	-116.1
	(52.08)	(181.3)	(188.4)	(168.6)	(147.6)	(127.1)
Interaction	-340.4***	4.573	-154.4*	-34.64	-136.1	-37.33
	(109.4)	(3.720)	(88.70)	(98.53)	(85.95)	(84.09)
Financial Treatment	56.11	292.1	98.15	408.3	41.34	100.2
	(94.05)	(575.7)	(297.1)	(275.0)	(307.0)	(286.4)
Financial Treatment*Female	-204.3**	-301.4	-336.4	-533.6**	-391.0	-220.2
	(87.30)	(498.1)	(275.7)	(244.7)	(260.7)	(202.4)
Interaction*Financial Treatment	46.93	-5.133	-19.28	-159.5	9.453	-18.18
	(145.8)	(11.37)	(132.5)	(121.5)	(139.7)	(115.4)
Interaction*Female	120.3	-7.119*	155.4*	43.71	85.45	81.56
	(86.35)	(3.798)	(88.52)	(78.22)	(69.00)	(54.87)
Interaction*Financial Treatment*Female	81.07	2.256	23.50	122.3	53.83	-20.96
	(135.8)	(9.779)	(122.9)	(105.5)	(114.7)	(82.05)
Observations	4101	4101	1992	1992	1992	1992
Clusters	152	152	115	115	115	115
Dep. Var. Mean	1195.9	1195.9	1195.9	1195.9	1195.9	1195.9
Dep. Var. SD	677.1	677.1	677.1	677.1	677.1	677.1

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 6: Effect of Financial Stakes by School Characteristics

	Predicted Raise Amount (USD)	
	(1) Student Withdrawal	(2) Teacher Turnover
Female	-66.98	56.03
	(93.20)	(209.0)
Interaction	-1272.9*	-140.2
	(717.8)	(602.7)
Financial Treatment	-434.7**	467.4
	(177.6)	(315.0)
Financial Treatment*Female	-40.20	-590.8**
	(147.7)	(252.9)
Interaction*Financial Treatment	4023.7***	-662.9
	(1270.8)	(681.0)
Interaction*Female	1057.7	43.86
	(746.4)	(445.4)
Interaction*Financial Treatment*Female	-1534.3	643.2
	(1118.7)	(543.6)
Observations	1961	1818
Clusters	152	128
Dep. Var. Mean	1193.3	1193.3
Dep. Var. SD	656.9	656.9

Notes: This table presents the relationship between the employee's performance evaluation score, the treatment status and school 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 school characteristic, with each column using a different characteristic. In column (1), *Interaction* is the average student withdrawal rate in the 5 years prior to the intervention (a suggestive measure of schooling market competitiveness). In column (2), it is the average teacher turnover rate in the 5 years prior to the intervention (a suggestive measure of labor market competitiveness). Standard errors are at the teacher level (the unit of randomization). $^{*}p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

Table 7: Effect of Information by Manager Type

	Predicted Raise Amount (USD)					
	(1) Male	(2) Age	(3) Avg. Bias	(4) Math Bias	(5) Jobs Bias	(6) Family Bias
Female	-126.5*	54.54	-314.8	-241.2	-310.2*	-72.88
	(66.62)	(382.6)	(190.1)	(185.6)	(173.1)	(177.1)
Interaction	-481.0***	-0.750	-90.07	-77.32	-119.6	10.58
	(92.48)	(8.025)	(86.26)	(76.88)	(78.34)	(85.25)
Observation Treatment	-119.1	-436.6	206.3	94.58	-169.5	246.9
	(78.35)	(365.8)	(338.4)	(202.3)	(235.1)	(318.1)
Observation Treatment*Female	165.9**	493.9	-78.30	42.03	278.0	-231.2
	(81.58)	(397.3)	(377.4)	(230.6)	(276.5)	(320.8)
Interaction*Observation Treatment	330.6**	7.586	-116.0	-72.14	59.84	-110.1
	(142.4)	(7.400)	(171.7)	(87.38)	(111.5)	(144.1)
Interaction*Female	302.6***	-2.545	118.0	89.51	126.4*	8.651
	(92.46)	(7.575)	(82.36)	(69.30)	(69.63)	(73.87)
Interaction*Observation Treatment*Female	-314.9*	-7.778	64.57	14.00	-107.0	109.1
	(181.8)	(7.925)	(180.8)	(98.15)	(131.6)	(139.1)
Observations	4101	4101	1992	1992	1992	1992
Clusters	152	152	115	115	115	115
Dep. Var. Mean	1195.9	1195.9	1195.9	1195.9	1195.9	1195.9
Dep. Var. SD	677.1	677.1	677.1	677.1	677.1	677.1

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). $^{*}p < 0.10$,** $p < 0.05$,*** $p < 0.01$.

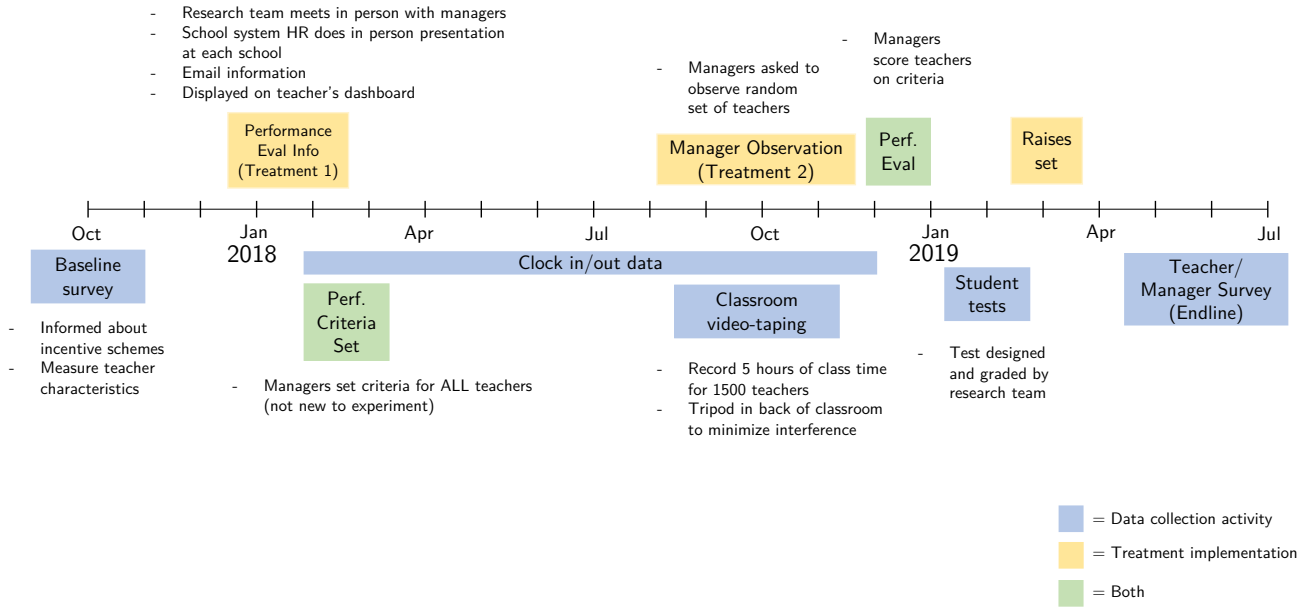
Table 8: Effect of Information by School Characteristics

	Predicted Raise Amount (USD)	
	(1) Student Withdrawal	(2) Teacher Turnover
Female	-244.8**	-391.8*
	(114.2)	(235.2)
Interaction	806.7	-548.5
	(1334.0)	(529.2)
Observation Treatment	-185.9	-124.3
	(116.1)	(213.5)
Observation Treatment*Female	225.9*	196.2
	(134.5)	(220.9)
Interaction*Observation Treatment	480.6	28.74
	(965.2)	(464.8)
Interaction*Female	681.5	452.7
	(905.9)	(482.8)
Interaction*Observation Treatment*Female	-490.1	-96.32
	(1263.4)	(478.3)
Observations	1961	1818
Clusters	152	128
Dep. Var. Mean	1193.3	1193.3
Dep. Var. SD	656.9	656.9

Notes: This table presents the relationship between the employee's performance evaluation score, the treatment status and school 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 school characteristic, with each column using a different characteristic. In column (1), *Interaction* is the average student withdrawal rate in the 5 years prior to the intervention (a suggestive measure of schooling market competitiveness). In column (2), it is the average teacher turnover rate in the 5 years prior to the intervention (a suggestive measure of labor market competitiveness). Standard errors are at the teacher level (the unit of randomization). $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

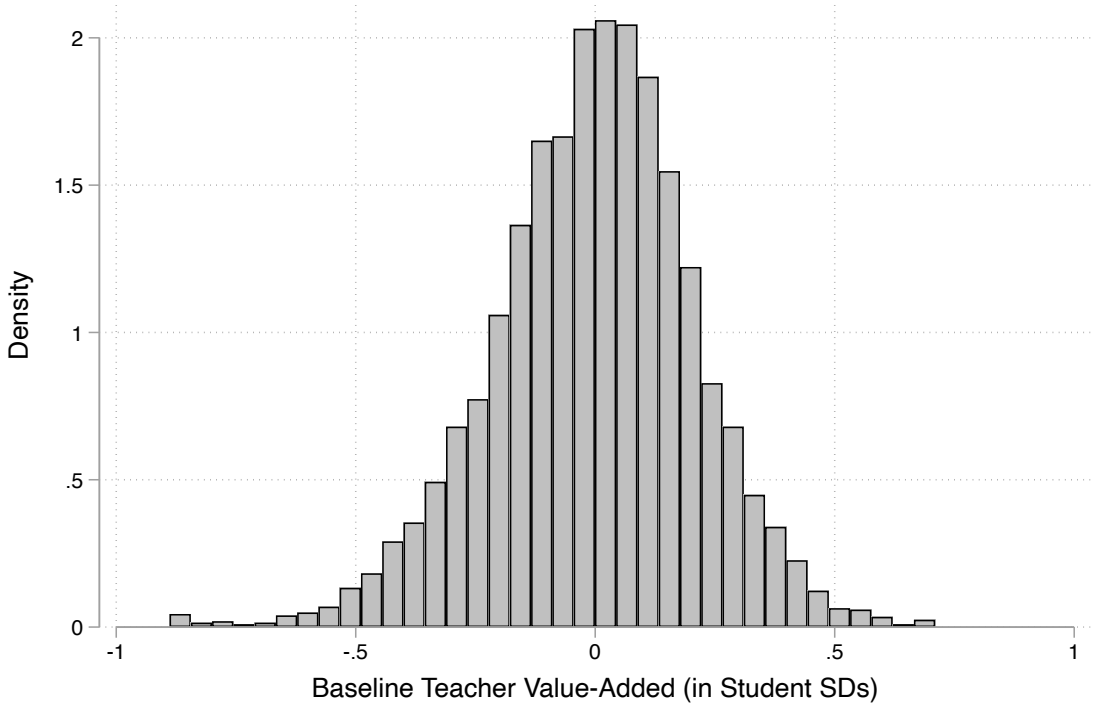
Appendix A - Supplementary Tables and Figures

Figure A1: Experimental Timeline



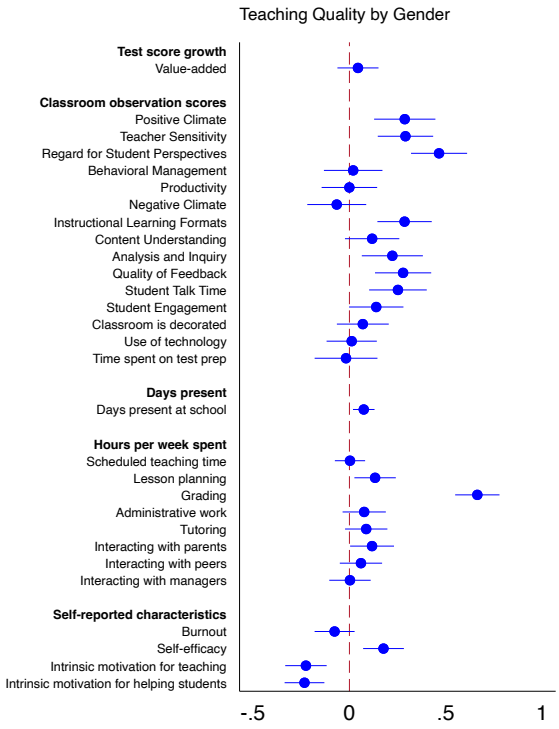
Notes: This figure shows the timeline of the experiment from October 2017 through July 2019 and includes treatment implementation activities and data collection activities/periods.

Figure A2: 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.

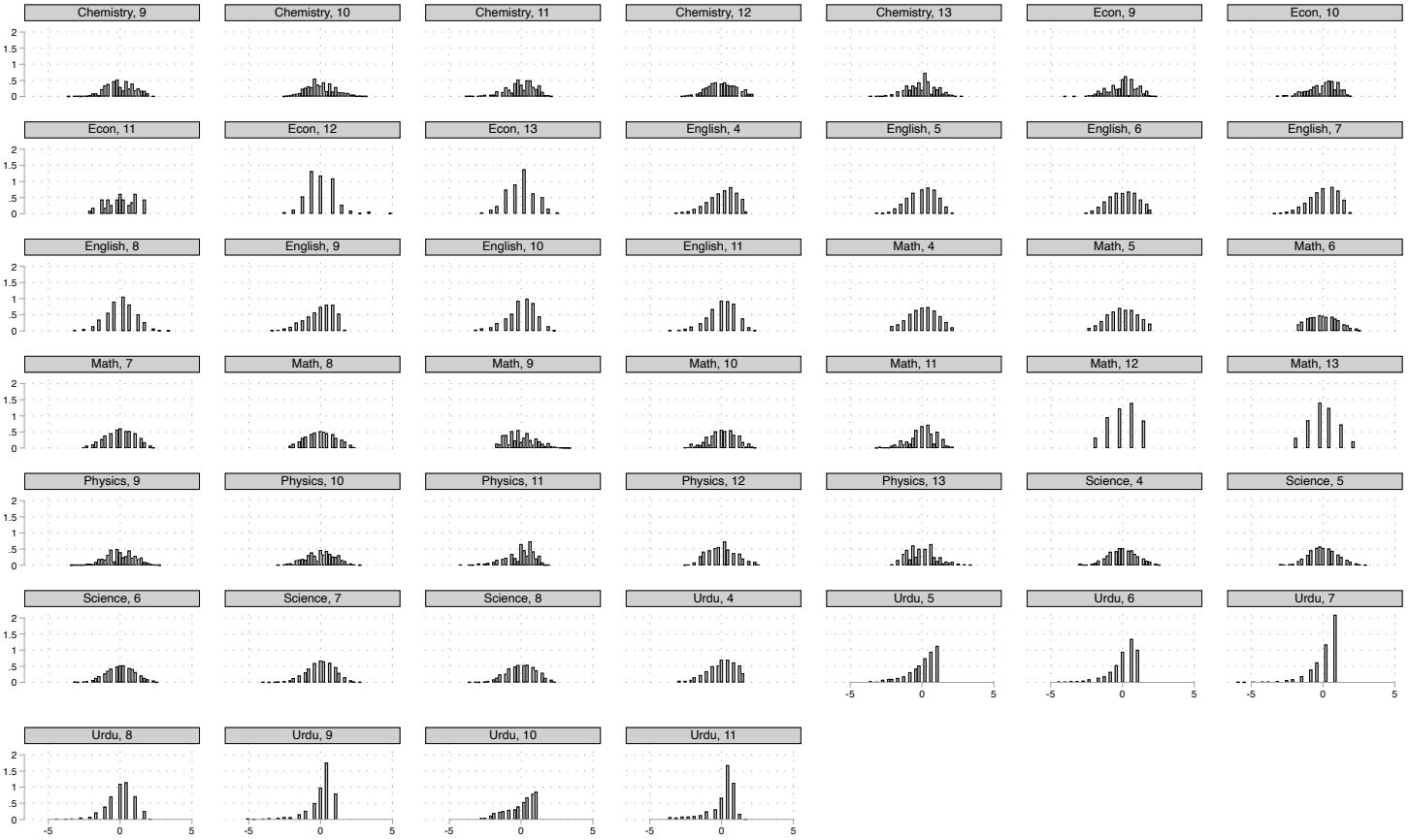
Figure A3: Teacher Effort and Productivity Measures by Gender



Notes: This figure shows the difference in various measures of teaching quality by teacher gender.

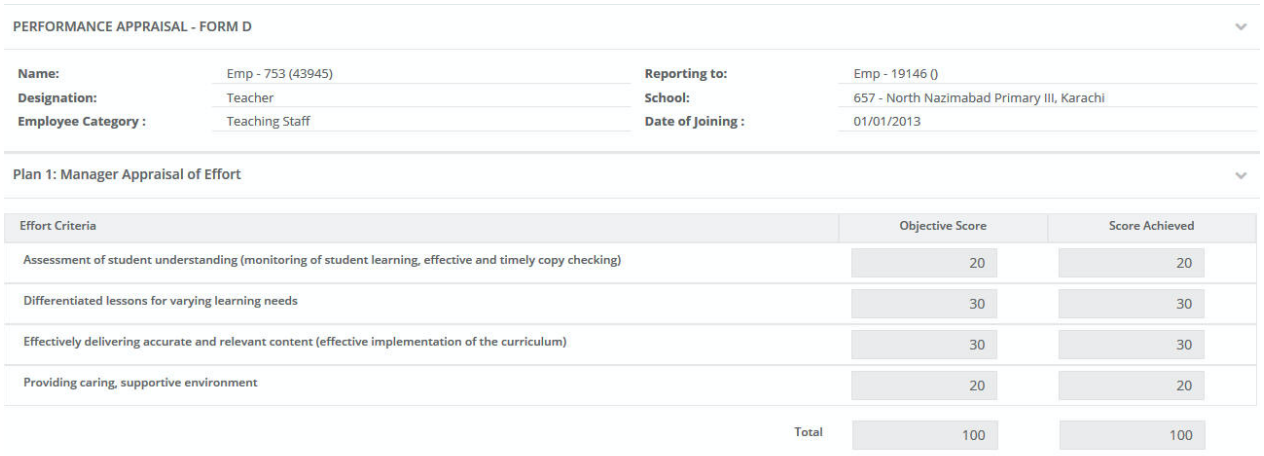
- The x-axis shows the magnitude of the coefficients in standard deviations and the y-axis shows the outcomes. A positive coefficient corresponds to a higher performance by female teachers.
- Data is from the endline survey, classroom observations and test scores.

Figure A4: 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 A5: 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

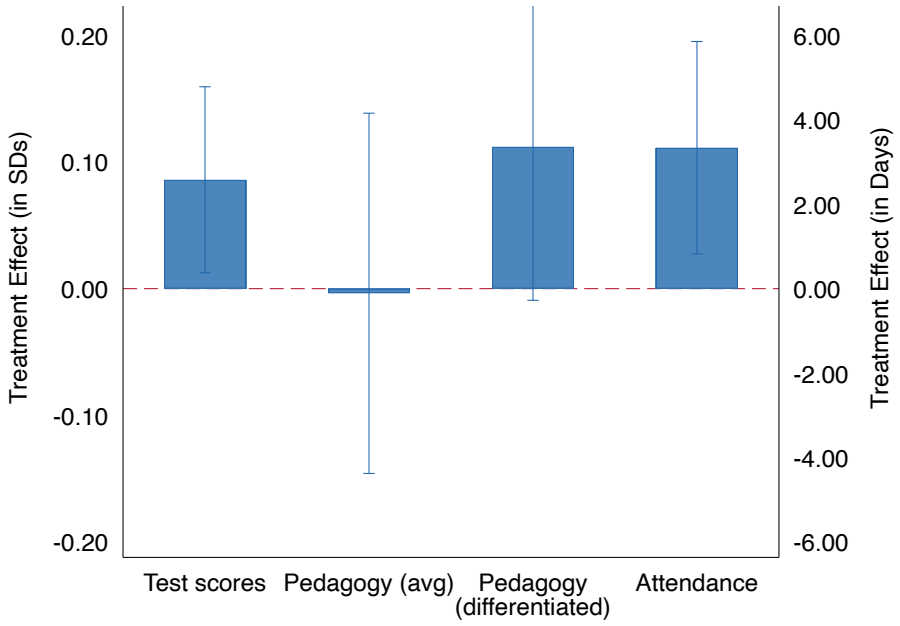
Notes: This figure shows an example set of performance criteria a teacher would have set in collaboration with their manager at the beginning of the year. This list of criteria was located on their employment portal, and available to access throughout the year. Managers could set individual criteria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

Table A1: Female Wage Gap using Bau and Das (2020)

	Monthly Wage (USD)				
	(1)	(2)	(3)	(4)	(5)
Female	-17.05***	-16.98***	-15.05***	-15.14***	-10.41***
	(1.967)	(1.983)	(1.777)	(1.819)	(2.030)
Value-added Math		0.307*	0.192	0.0744	4.529*
		(0.172)	(0.163)	(0.160)	(2.525)
Value-added English		2.243	0.508	0.00386	3.716
		(3.167)	(2.895)	(2.903)	(2.602)
Value-added Urdu		-2.517*	-1.817	-0.970	-4.460*
		(1.450)	(1.323)	(1.253)	(2.608)
Days absent during last month		-0.490**	-0.492***	-0.401**	0.0208
		(0.194)	(0.182)	(0.190)	(0.203)
Some post-HS training			7.840***	7.155***	58.68***
			(1.539)	(1.470)	(9.294)
BA or greater			10.21***	9.306***	-8.336***
			(1.524)	(1.610)	(2.056)
Constant	37.83***	38.41***	31.76***	32.16***	78.40***
	(1.907)	(1.978)	(1.638)	(1.627)	(9.368)
Gender Wage Gap	-46%	-46%	-41%	-41%	-7%
Mean (male teachers)	36.69	36.69	36.69	36.69	142.10
Clusters	975	975	975	975	1528
School Sample	Private	Private	Private	Private	Public
Controls					
Value-Added/Attendance		X	X	X	X
Credentials			X	X	X
School Fixed Effects				X	X

Notes: This table shows the relationship between teacher wage, gender, measures of teacher productivity and school type using data from Bau and Das (2020). $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

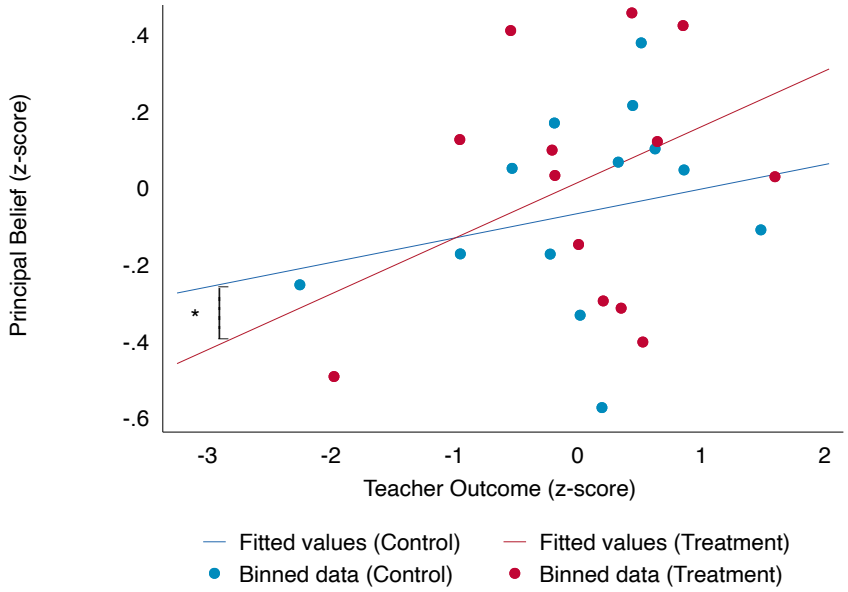
Figure A6: Effect of Financial Stakes on Student Outcomes and Teacher Behavior



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.

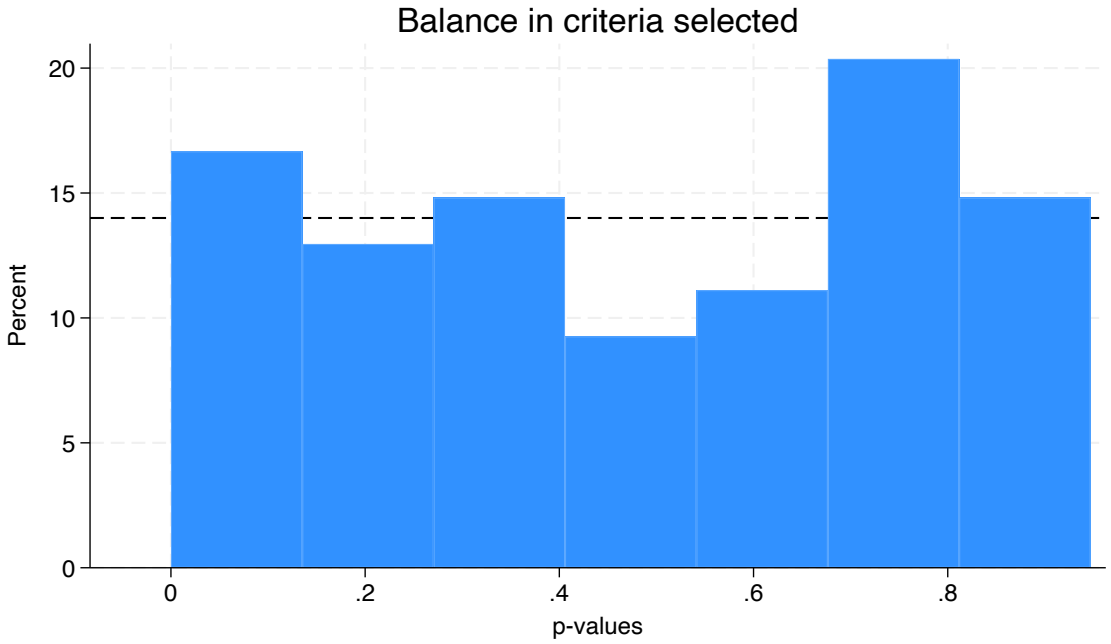
Figure A7: Effect of Observation Treatment on Principal Beliefs



Notes: This figure shows the relationship between our measured teacher quality and manager's prediction of quality on that dimension 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. The measures of productivity include: attendance, value-added, classroom climate and use of higher order thinking skills.
- Observations are at teacher-outcome level. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

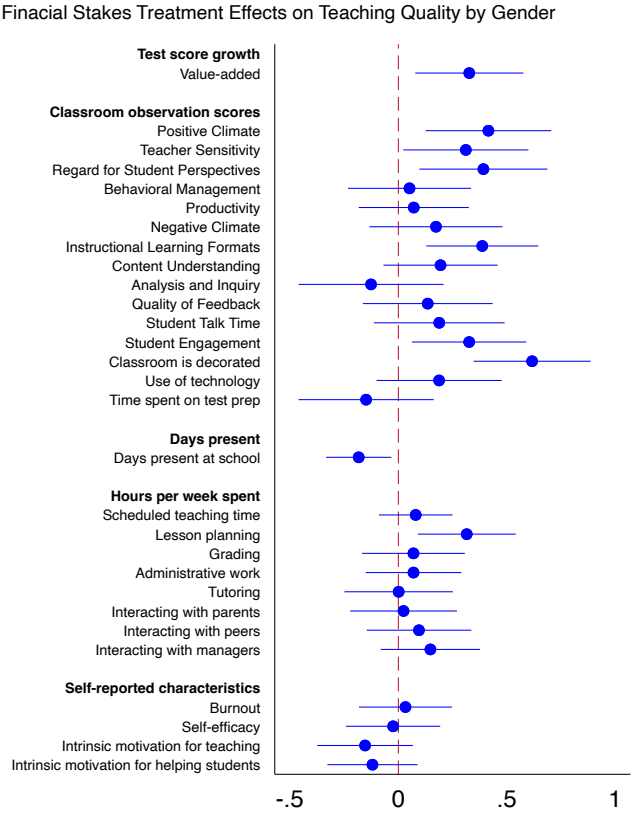
Figure A8: Performance Evaluation Criteria Weight by Treatment



Notes: This figure shows the distribution of p-values for a series of regressions testing whether teachers under the financial stakes treatment were evaluated using different types of performance criteria than those in the control.

- The figure shows a histogram of p-values from 54 regressions. The dotted red line shows the expected p-value distribution for an infinite number of regressions where the true population effect is zero.
- The regressions conducted were: *For i = 1,..., 27: (Any evaluation points for category i)*$_j = \beta_0 + \beta_1 FinancialStakes_j + \epsilon_j$
  and
  (% of evaluation points for category i)$_j = \beta_0 + \beta_1 FinancialStakes_j + \epsilon_j$
  where i indexes the possible categories of evaluation such as accuracy of content, communication skills, contribution to extracurriculars, engagement in professional development, etc for teacher j.
- Data on the text of the evaluation criteria comes from administrative employment records provided by the school system. The text responses were then categorized into 27 groups by the research team.

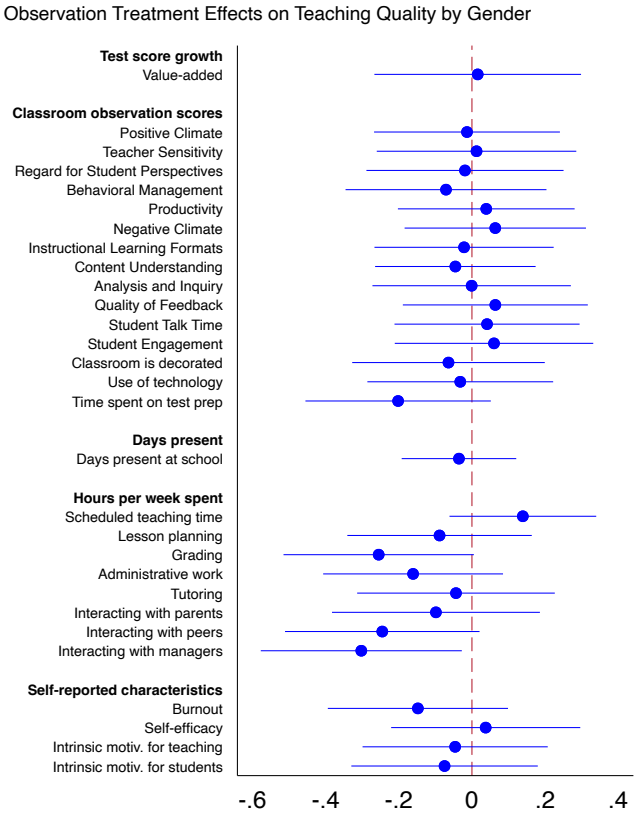
Figure A9: Heterogeneous Effect of Financial Stakes on Teaching Quality by Gender



Notes: This figure shows the heterogeneous effect by teacher gender of financial stakes versus no financial stakes of performance evaluation (pooling flat raises and test-score based incentive raises) on various measures of teaching quality.

- The x-axis shows the magnitude of the coefficients in standard deviations and the y-axis shows the outcomes. A positive coefficient corresponds to a larger treatment effect for female teachers than male teachers.
- Data is from the endline survey, classroom observations and test scores.

Figure A10: Heterogeneous Effect of Observation Treatment on Teaching Quality by Gender

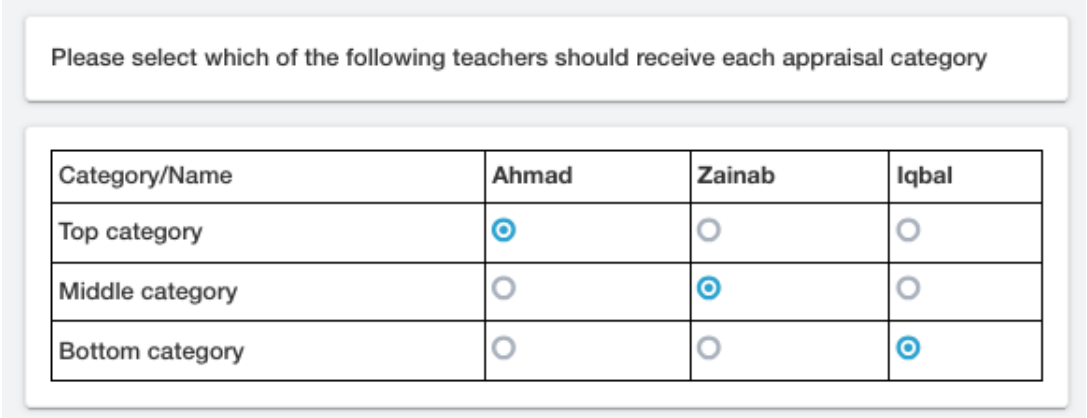


Notes: This figure shows the heterogeneous effect by teacher gender of the observation treatment on various measures of teaching quality.

- The x-axis shows the magnitude of the coefficients in standard deviations and the y-axis shows the outcomes. A positive coefficient corresponds to a larger treatment effect for female teachers than male teachers.
- Data is from the endline survey, classroom observations and test scores.

Figure A11: Vignette Teacher Profile Example

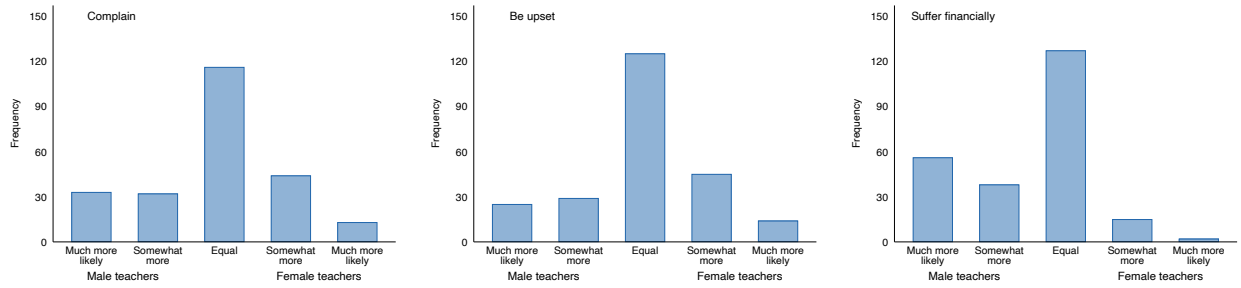
Name	Ahmad	Zainab	Iqbal
Teach	Class 5 Urdu	Class 1 Urdu	Class 1 Math
Test score growth	average	above average	below average
Classroom environment	organized and supportive of learning	disorganized and noisy	organized and supportive of learning
Days of leave	0 days, much less than average	7 days, about average	10 days, more than average
Classroom observation	You have observed the teacher frequently, so you are confident in your assessment of them.	You have not observed the teacher this year, so you are uncertain about their performance.	You have observed the teacher frequently, so you are confident in your assessment of them.
Plans for next year	Staying at your school	Transferring to another school	Transferring to another school
Years working with teacher	1	8	1
Spouse's job	Does not work	Doctor	Teacher



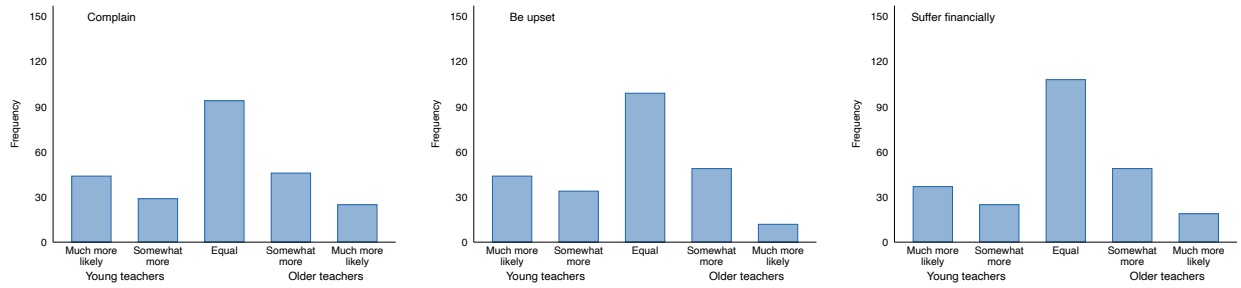
Notes: This figure shows an example of the choice managers would see in the vignette survey. Managers would be shown three teachers with descriptions of their performance and characteristics. They would then be asked to rank the three teacher's performance.

Figure A12: Managers' Beliefs about Consequence of Low Raise by Teacher Demographics

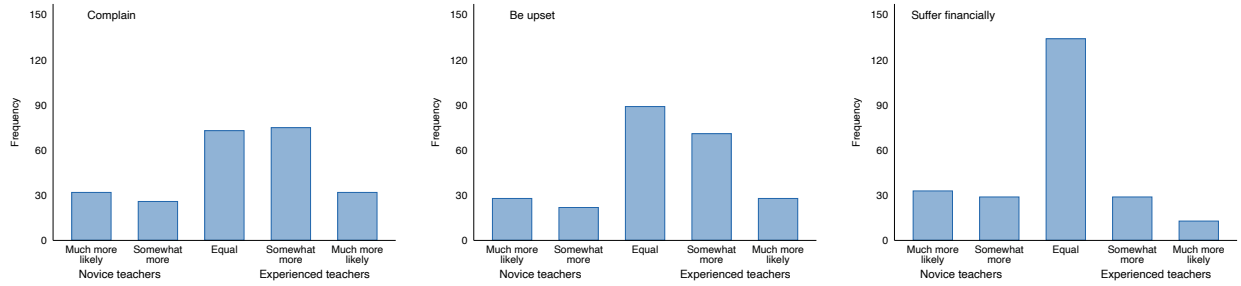
Panel A: Who is more likely to _____ if they receive a low raise?



Panel B: Who is more likely to _____ if they receive a low raise?



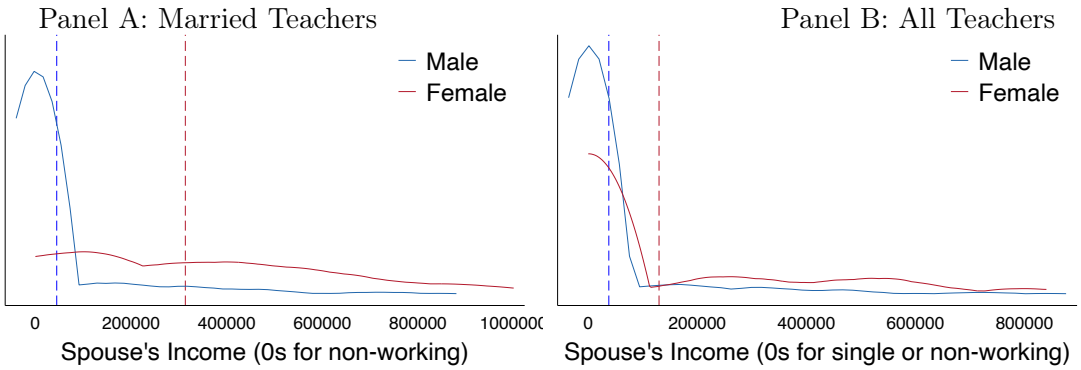
Panel C: Who is more likely to _____ if they receive a low raise?



Notes: This figure shows the distribution of survey responses about which types of employees are more likely to respond negatively to a low raise.

- Each of the nine figures shows the distribution of responses a separate survey question. From left to right the figures show the response the question "Who is more likely to [complain/be upset/suffer financially] if they receive a low raise?".
- Panel A shows the responses when the choices vary from "Male teachers are much more likely" to "Female teachers are much more likely".
- Panel B shows the responses when the choices vary from "Young teachers are much more likely" to "Older teachers are much more likely".
- Panel C shows the responses when the choices vary from "Novice teachers are much more likely" to "Experienced teachers are much more likely".
- Data is from the manager vignette survey conducted with 239 principals in spring 2023.

Figure A13: Spousal Income by Teacher Gender



Notes: This figure shows the distribution of spouse's income by employee gender.

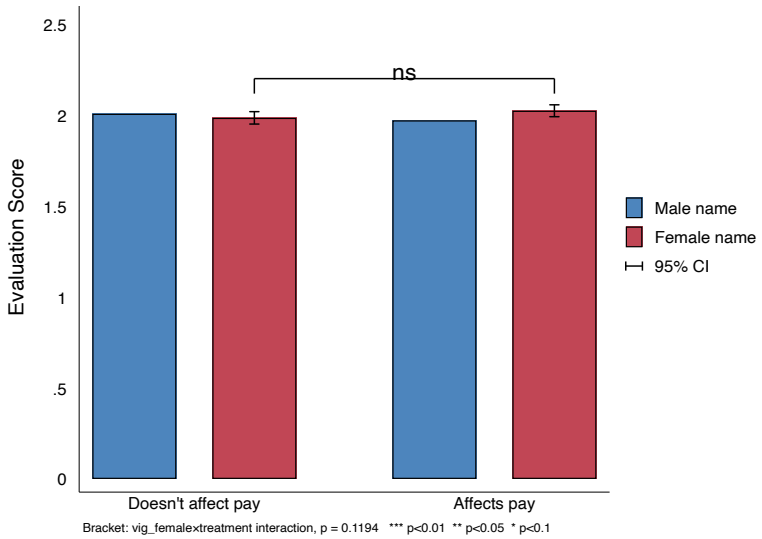
- The data comes from the 2018 Pakistan Social and Living Standards Measurement survey (PSLM) restricting to the sample which listed occupation as primary or secondary school teacher.
- The blue and red solid line show the density function of spouse's income for male versus female teachers, respectively. The blue and red dotted line show the mean spouse income.
- Panel A restricts to married teachers. Panel B includes all teachers and counts unmarried teachers as having a zero for spouse's income.

Table A2: Female Wage Gap using Experimental Sample

	Monthly Wage (USD)			
	(1)	(2)	(3)	(4)
Female	-8055.0***	-4859.7***	-2637.1***	-2824.1***
	(1131.3)	(1123.0)	(969.7)	(1006.6)
Observations	720	720	720	720
Gender Wage Gap	-31%	-19%	-10%	-11%
Mean (male teachers)	26229.03	26229.03	26229.03	26229.03
School Sample				
Controls				
Value-Added/Video Scores		X	X	X
Experience			X	X
School Fixed Effects				X
Grade/Subject Fixed Effects				X

Notes: This table shows the relationship between teacher wage, gender, measures of teacher productivity and school type using data from main experimental sample. $^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$.

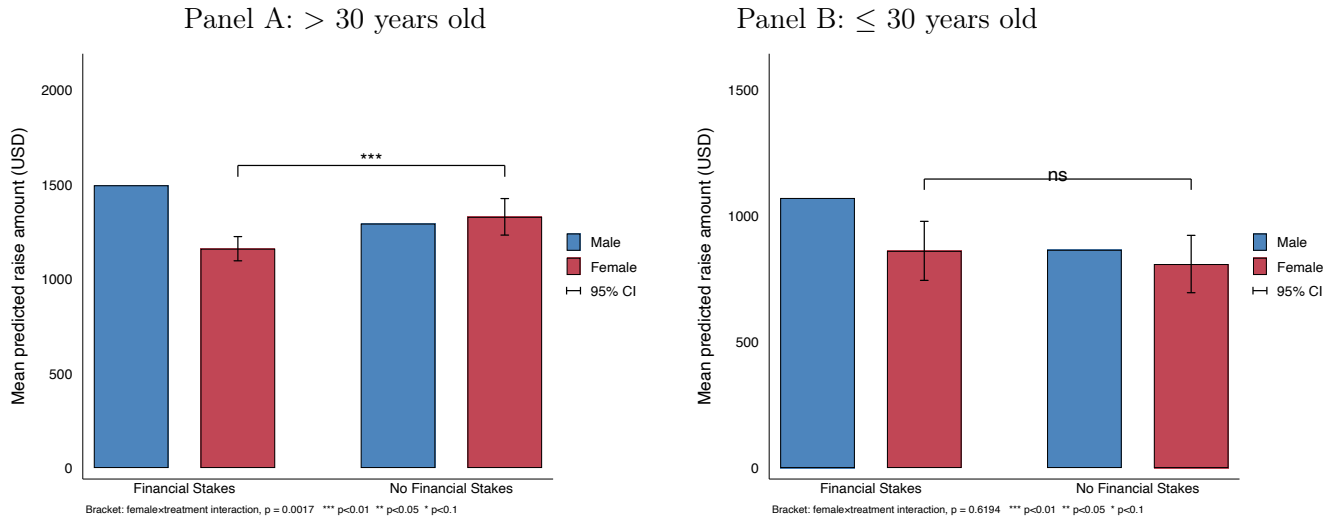
Figure A14: Effect of financial stakes on evaluation score by profile gender



Notes: This figure shows the effect of financial stakes on the evaluation score by profile gender from the hypothetical vignette survey conducted with principals.

- The y axis is the average evaluation score given by survey respondents. The first two bars show the average rating by profile gender for respondents who were told the score would not affect pay. The second two bars show the average rating profile gender for respondents who were told the score would affect pay. The bracket above the bars shows the statistical significance of the interaction between financial stakes and profile gender.
- Data is from the manager vignette survey conducted with 239 principals in spring 2023.
- Standard errors are clustered at the manager level (the unit of randomization). $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Figure A15: Effect of Financial Stakes Treatment on Evaluation Score by Gender and Age



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.
- Panel A restricts to teachers who are older than 30 years old. Panel B restricts to teachers who are less than or equal to 30 years old.
- 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. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table A3: 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 A4: 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 A5: 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 A6: 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 A7: Baseline Covariates

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

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

Table A8: 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.

Table A9: Raise Percent by Treatment and Gender, PDS Lasso Controls

	Predicted Raise Percent			
	(1)	(2)	(3)	(4)
Female	0.00234**	0.00284	0.00124	0.00310
	(0.00107)	(0.00184)	(0.00138)	(0.00225)
Financial stakes		-0.00175		0.00213
		(0.00194)		(0.00253)
Financial stakes*Female		-0.000878		-0.00297
		(0.00217)		(0.00279)
Observation Treatment			-0.00143	0.00350
			(0.00167)	(0.00265)
Observation Treatment*Female			0.00230	-0.000115
			(0.00186)	(0.00271)
Financial stakes*Observation Treatment				-0.00794**
				(0.00330)
Financial stakes*Observation Treatment*Female				0.00391
				(0.00362)
Observations	2806	2806	2806	2806
Clusters	167	167	167	167
Dep. Var. Mean	0.0661	0.0661	0.0661	0.0661
Dep. Var. SD	0.0204	0.0204	0.0204	0.0204

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 percentage 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. *Observation 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 strata and double lasso selected controls from the set of: 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 A10: Raise Percent by Treatment and Gender, Rank-Ordered Logit

	Predicted Raise Percent			
	(1)	(2)	(3)	(4)
Female	0.136**	0.194**	0.0259	0.165
	(0.0538)	(0.0958)	(0.0734)	(0.120)
Financial stakes		0		0
		(.)		(.)
Financial stakes*Female		-0.0903		-0.217
		(0.109)		(0.142)
Observation Treatment			-0.156*	0.0900
			(0.0798)	(0.129)
Observation Treatment*Female			0.231***	0.0875
			(0.0857)	(0.136)
Financial stakes*Observation Treatment				-0.385**
				(0.154)
Financial stakes*Observation Treatment*Female				0.224
				(0.166)
Observations	2806	2806	2806	2806
Clusters	167	167	167	167
Dep. Var. Mean	0.0661	0.0661	0.0661	0.0661
Dep. Var. SD	0.0204	0.0204	0.0204	0.0204

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

- The dependent variable is the employee's evaluation score converted into the associated raise percentage 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. *Observation 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 strata and double lasso selected controls from the set of: 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 A11: 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)	0.166***	0.192***	0.231**	0.136	-0.0484	0.0579
	(0.0434)	(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 A12: Correlation between Predicted Raise and Teacher Quality by Treatment

	Predicted Raise Amount (USD)				
	(1) Value-added	(2) Observation Score	(3) Days Present	(4) Total hours	(5) Hours per day
Interaction	35.58	22.39	648.1***	420.5***	139.3
	(30.79)	(59.87)	(130.6)	(139.9)	(161.7)
Financial Treatment	-91.80*	-88.68	-1.905	-79.23*	-157.0
	(50.39)	(79.46)	(120.5)	(45.25)	(123.4)
Interaction*Financial Treatment	-40.88	-95.75	-109.5	-182.5	-109.6
	(33.20)	(68.35)	(150.9)	(171.0)	(184.7)
Observations	2484	989	2140	2140	2140
Clusters	214	125	240	240	240
Dep. Var. Mean	1193.3	1193.3	1193.3	1193.3	1193.3
Dep. Var. SD	656.9	656.9	656.9	656.9	656.9

Notes: This table presents the relationship between the employee's predicted raise (evaluation score group*base salary), measured quality, treatment status and gender. The dependent variable is the employee's evaluation score converted into the associated raise value in USD for that score. *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 measure of teacher quality (listed at the top of the column). In column (1), *Interaction* is value-added z-score (2), it is classroom observation z-score. In column (3), it is days present. Column (4) is total hours present in the school and column (5) is mean hours present per day in school. Standard errors are clustered at the school level (the unit of randomization). $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table A13: Correlation between Predicted Raise and Teacher Quality by Treatment and Gender

	Predicted Raise Amount (USD)				
	(1) Value-added	(2) Observation Score	(3) Days Present	(4) Total hours	(5) Hours per day
Female	-26.29	138.4	-361.4	5.618	625.0**
	(100.2)	(127.0)	(302.8)	(54.55)	(245.4)
Interaction	-34.87	-186.5*	221.9	-321.8	-676.2*
	(55.40)	(95.26)	(367.0)	(353.1)	(394.1)
Financial Treatment	97.90	282.0*	-936.9**	26.56	504.0*
	(123.7)	(145.5)	(367.6)	(79.96)	(291.2)
Financial Treatment*Female	-229.4**	-436.3***	1071.8***	-137.1*	-782.3***
	(113.8)	(155.8)	(378.9)	(70.24)	(276.8)
Interaction*Financial Treatment	109.4	132.7	1211.8**	731.3*	744.4*
	(71.07)	(115.3)	(475.7)	(432.6)	(438.2)
Interaction*Female	75.49	238.4**	506.5	878.7**	956.9**
	(59.93)	(101.1)	(391.2)	(346.8)	(381.2)
Interaction*Financial Treatment*Female	-162.2**	-235.8*	-1539.1***	-1082.6**	-987.0**
	(76.01)	(122.6)	(488.5)	(434.9)	(433.2)
Observations	2484	989	2140	2140	2140
Clusters	214	125	240	240	240
Dep. Var. Mean	1193.3	1193.3	1193.3	1193.3	1193.3
Dep. Var. SD	656.9	656.9	656.9	656.9	656.9

Notes: This table presents the relationship between the employee's predicted raise (evaluation score group*base salary), measured quality, treatment status and gender. The dependent variable is the employee's evaluation score converted into the associated raise value in USD for that score. *Female* is a dummy for whether the employee is female. *Financial Treatment* is a dummy which is 1 if the teacher's school was assigned to have their evaluation determine their raise or 0 if their evaluation was just for feedback purposes. *Interaction* is a measure of teacher quality (listed at the top of the column). In column (1), *Interaction* is value-added z-score (2), it is classroom observation z-score. In column (3), it is days present. Column (4) is total hours present in the school and column (5) is mean hours present per day in school. Standard errors are clustered at the school level (the unit of randomization). $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table A14: Performance Evaluation Criteria Text Attributes by Treatment Status

	(1) # of characters	(2) Round numbers	(3) Score SD (within school)	(4) Score SD (within teacher)
Financial stakes	3.211	-0.0126	-0.00870	-0.00193
	(5.090)	(0.0233)	(0.00567)	(0.00399)
Observations	17294	54198	54779	54339
Clusters	255	278	279	278
Dep. Var. Mean	110.7	0.313	0.112	0.110
Dep. Var. SD	57.95	0.464	0.0449	0.0442

Notes: This table shows different measures of manager performance evaluation effort and tests whether this is differential by treatment status. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table A15: Vignette Survey Questions

Section A:
Select whether you agree or disagree the following will happen in response to this teacher receiving a [low/middle/high] rating? 5 point scale from Strongly disagree to Strongly agree

1. Teacher would complain to me
2. Teacher would feel happy
3. Teacher may begin looking for another job
4. Teacher may be less willing to help with extra tasks
5. Teacher would feel disappointed
6. Teacher may volunteer to assist with extra duties
7. I would feel bad for the teacher
8. I would feel good about the decision
9. Teacher would suffer financial hardship

Section B:
1. Which types of teachers are more likely to complain about a low raise?
5 point scale from: 1a. *Young teachers are much more likely* to *Older teachers are much more likely*
1b. *Male teachers are much more likely* to *Female teachers are much more likely*
1c. *Inexperienced teachers are much more likely* to *Experienced teachers are much more likely*
2. Which types of teachers are more likely to be upset about a low raise?
5 point scale from: 2a. *Young teachers are much more likely* to *Older teachers are much more likely*
2b. *Male teachers are much more likely* to *Female teachers are much more likely*
2c. *Inexperienced teachers are much more likely* to *Experienced teachers are much more likely*
3. Which types of teachers are more likely to suffer financial hardship due to a low raise?
5 point scale from: 3a. *Young teachers are much more likely* to *Older teachers are much more likely*
3b. *Male teachers are much more likely* to *Female teachers are much more likely*
3c. *Inexperienced teachers are much more likely* to *Experienced teachers are much more likely*

Notes: This figure shows the survey questions managers were asked to evaluate how an employee would react to a given evaluation score. In section A, managers were shown one of the teacher vignette profiles and told the teacher would receive a [low/middle/high] rating. Then they were asked how likely each of the following responses by the teacher would be. In section B, managers were asked to respond to general statements about what types of teachers were more or less likely to complain/be upset/suffer financial hardship if given a low rating.

Table A16: Vignette Attribute Values

Attribute	**Possible Values**
Teacher name	*Female names:* Ayesha, Bushra, Fatima, Hafsa, Maliha Maryam, Sarah, Soraya, Tahira, Yasmin, Zahida, Zainab *Male names:* Ahmad, Abdul, Akmal, Farhan, Iqbal, Kamran, Muhammad Nadeem, Omar, Shahid, Tariq, Usman
Subject and grade taught	Class 1 Urdu Class 1 Math Class 5 Urdu Class 5 Math
Test score growth	below average average above average
Classroom environment	organized and supportive of learning disorganized and noisy
Days of leave	0 days, much less than average 7 days, about average 10 days, more than average
Classroom observation	You have observed the teacher frequently, so you are confident in your assessment of them. You have not observed the teacher this year, so you are uncertain about their performance.
Plans for next year	Staying at your school Transferring to another school
Years working with teacher	1 8
Spouse's job	Does not work Teacher Doctor

Notes: This table shows the possible values for each of the vignette dimensions shown to managers.

Table A17: Turnover by Appraisal Score and Gender - IV

	(1) Turnover
Female	1.264
	(1.449)
Appraisal rating (Instrumented)	0.309
	(0.497)
Female*Appraisal rating (Instrumented)	-0.467
	(0.547)
Observations	12971
Clusters	385
Dep. Var. Mean	0.296
Dep. Var. SD	0.456
First stage F stat.	18.09

Notes: This table presents the relationship between turnover and employee's appraisal score by gender. Appraisal rating is instrumented by the mean appraisal rating of new entrants to the school. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table A18: VA by Appraisal Score and Gender - IV

	Value-Added	
	(1) Same year	(2) Next year
Female	0.0165	0.790
	(1.158)	(1.542)
Appraisal rating (Instrumented)	0.120	0.344
	(0.387)	(0.500)
Female*Appraisal rating (Instrumented)	-0.00215	-0.292
	(0.438)	(0.583)
Observations	7451	5998
Clusters	294	286
Dep. Var. Mean	0.296	0.296
Dep. Var. SD	0.456	0.456
First stage F stat.	18.09	18.09

Notes: This table presents the relationship between value-added (same and following year) and employee's appraisal score by gender. Appraisal rating is instrumented by the mean appraisal rating of new entrants to the school. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table A19: Teacher Salary by Gender, Marital Status and School Type

	(1) Log Salary (USD)
Female	-0.113
	(0.0779)
Married	-0.00475
	(0.0369)
Female * Married	0.222***
	(0.0694)
Female * Private School	-0.179*
	(0.0946)
Married*Private School	0.0963
	(0.0697)
Female*Married*Private School	-0.209**
	(0.100)
Observations	3067

Notes: This table shows the relationship between teacher salary, gender, marital status and school type using data from Bau and Das (2020). $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

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:

$$y_{ijkcst} = \beta_0 + \sum_s \beta_s y_{ijkcs,t\text{-}1} \mathbb{1}[\textit{subject-grade} = s] + \sum_s \alpha_s y_{ijkcs,t\text{-}2} \mathbb{1}[\textit{subject-grade} = s] \quad (12)$$
$$+ \sum_s \gamma_s \bar{y}_{\text{-}ijkcs,t\text{-}1} \mathbb{1}[\textit{subject-grade} = s] + \chi_{st} + \psi_k + v_{ijkcst}$$
$$\textit{where } v_{ijkcst} = \mu_j + \theta_{ct} + \epsilon_{ijkcst} \quad (13)$$

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 A2 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.